

Holding Accountability to Account:

How Scholarship and
Experience in Other Fields
Inform Exploration of
Performance Incentives
in Education

Richard Rothstein

Prepared for *Performance Incentives:
Their Growing Impact on American K-12 Education*
in Nashville, Tennessee on February 28, 2008

Working Paper 2008-04
February 2008

LED BY



VANDERBILT
PEABODY COLLEGE

IN COOPERATION WITH:



 Mizzou
University of Missouri - Columbia

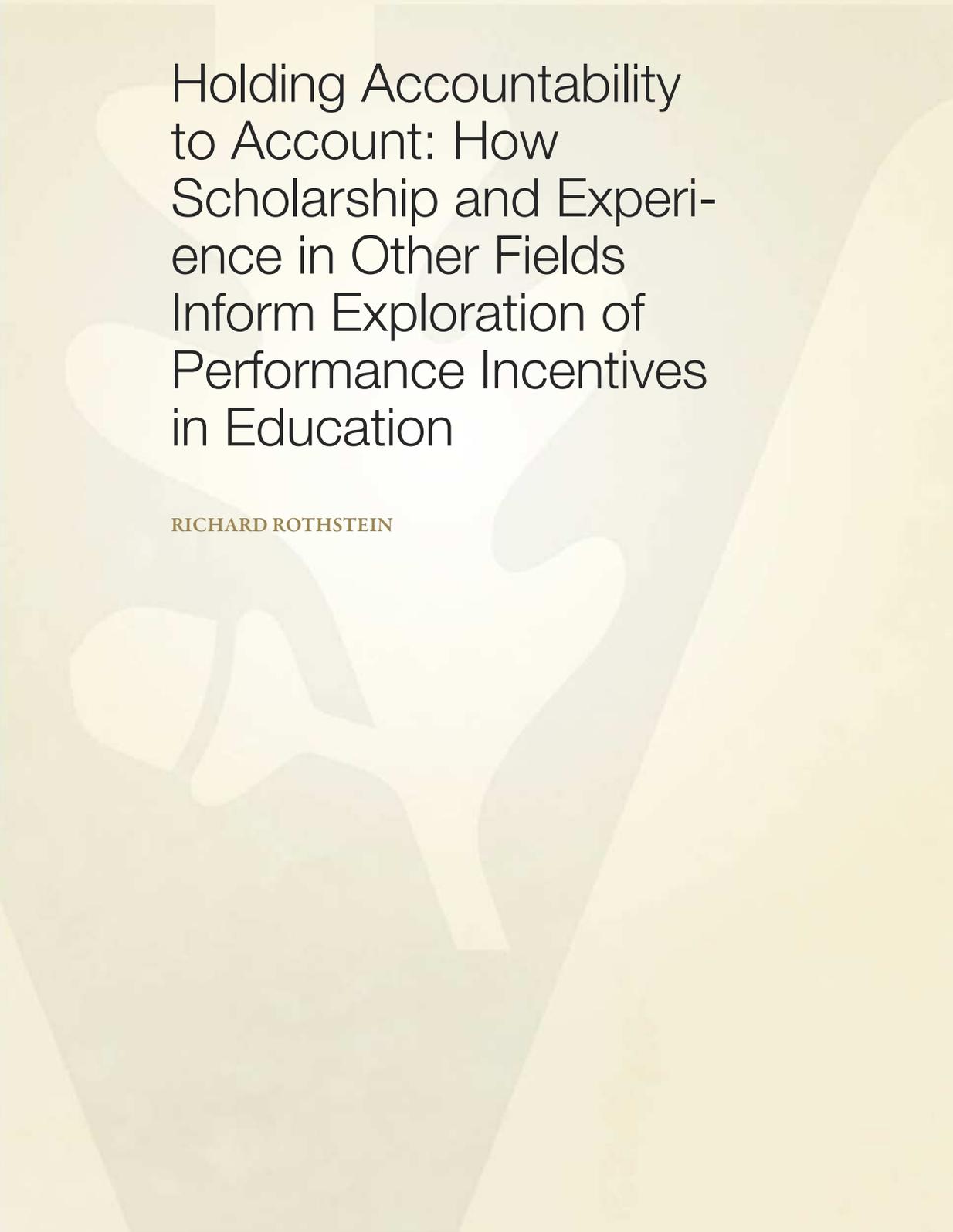
THE NATIONAL CENTER ON PERFORMANCE INCENTIVES

(NCPI) is charged by the federal government with exercising leadership on performance incentives in education. Established in 2006 through a major research and development grant from the United States Department of Education's Institute of Education Sciences (IES), NCPI conducts scientific, comprehensive, and independent studies on the individual and institutional effects of performance incentives in education. A signature activity of the center is the conduct of two randomized field trials offering student achievement-related bonuses to teachers. The Center is committed to air and rigorous research in an effort to provide the field of education with reliable knowledge to guide policy and practice.

The Center is housed in the Learning Sciences Institute on the campus of Vanderbilt University's Peabody College. The Center's management under the Learning Sciences Institute, along with the National Center on School Choice, makes Vanderbilt the only higher education institution to house two federal research and development centers supported by the Institute of Education Services.

This working paper was supported by the National Center on Performance Incentives, which is funded by the United States Department of Education's Institute of Education Sciences (R30SA06034). This is a draft version of a paper that will be presented at a national conference, *Performance Incentives; Their Growing Impact on American K-12 Education*, in Nashville, Tennessee on February 28-29, 2008. The author acknowledges additional support from the Campaign for Educational Equity, Teachers College, Columbia University. The views expressed in this paper do not necessarily reflect those of sponsoring agencies or individuals acknowledged. Any errors remain the sole responsibility of the author.

Please visit www.performanceincentives.org to learn more about our program of research and recent publications.



Holding Accountability
to Account: How
Scholarship and Experi-
ence in Other Fields
Inform Exploration of
Performance Incentives
in Education

RICHARD ROTHSTEIN

Contents:

Introduction.....	3
Part I: Mismeasurement of Outputs.....	9
Goal Distortion.....	9
Performance Thresholds.....	24
Part II: Mismeasurement of Inputs.....	30
Defining Subgroups, or Risk Adjustment.....	30
'Cream-Skimming'.....	40
Part III: Untrustworthy Statistics.....	46
Data Reliability.....	46
Sampling Corruption.....	49
Other Gaming.....	52
Part IV: The Private Sector.....	59
Part V: Intrinsic Motivation.....	72
Conclusion.....	78
Postscript.....	81
Bibliography.....	82
Acknowledgements.....	95
Endnotes (Citations to Bibliography).....	97

Introduction

In 1935, a 19 year-old political science major at the University of Chicago interviewed Milwaukee administrators for a term paper. He was puzzled that school board and public works officials could not agree on whether to invest marginal parks funds in play-supervision or in physical maintenance. He concluded that rational decision making was impossible because officials weighted parks goals differently; school board members thought mostly of recreational opportunities, while public works administrators thought mostly of green space to reduce density.

The next year, the director of the International City Managers' Association hired the young graduate as his 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 had difficult-to-define objectives. Most services have multiple purposes and even if precise definition were possible, evaluation would require difficult judgments to weight these purposes. Also, it was never possible to quantify whether outcome differences between cities were attributable to differences in effort and competence, 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 managers' 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 organizational behavior is characterized by "bounded rationality": weighing measurable costs and benefits does "not even remotely describe the processes that human beings use for making decisions in complex situations."¹

Undaunted, policymakers have recently devised quantitative incentive systems to maximize public service efficiency. In Great Britain, Margaret Thatcher attempted to rationalize public enterprises: where they could not be privatized, her government hoped to regulate them, using rewards and sanctions for quantifiable outcomes. Tony Blair accelerated these efforts, while in the United States, the Clinton Administration proposed to similarly "reinvent government." The Government Performance Results Act of 1993 (GPRA) demanded a shift in attention from processes towards measurable outcomes.

These efforts took little account of Herbert Simon's insights and ignored warnings of the great methodologist, Donald T. Campbell, who concluded that attempts to reward institutional behavior should account for actors who behaved differently when they were being measured.

Social scientists have long been aware of possible Hawthorne effects, so named because factory workers apparently behaved differently when being studied. Almost a Heisenberg uncertainty principle for human behavior, the Hawthorne effect suggests it is difficult to get human beings to 'stand still' long enough for their activity to be measured. At the Hawthorne Electric factory in the 1920s, workers simply stepped up efforts when they were studied (both when their work areas were made brighter, and dimmer), perhaps to make themselves look better to social scientists.

But Hawthorne workers had no personal stake in the research findings, no financial or security incentives to trick observers into believing performance was better than, in fact, it typically was. Donald Campbell, however, was concerned not with social science research generally but with accountability and control systems specifically. In these, possibilities of rewards or punishments create incentives for agents to appear more competent, even employing

deception and fraud to improve the impression. In 1979, Campbell framed what he called a '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

The law summarized efforts to use quantitative output indicators not only in education but in business, health care, welfare policy, human capital development, criminal justice, and public administration.

Simon and Campbell defined two shoals on which public accountability policy has foundered: that public goals are too complex to reduce to simple quantifiable measures; and attempts to do so corrupt public service.

As policy makers in education now grapple with re-authorizing *No Child Left Behind*, they confront examples of such corruption and are, fruitlessly it seems, attempting to 'fix' the law accordingly. They face three obstacles:

1) Conventional definitions and measurements of educational *outputs* are so oversimplified that they cannot support valid accountability or performance incentive systems. *Goal distortion* results, including re-allocation of resources to tested curricular areas from non-tested areas (like art, music, science, social studies or physical education); and increased focus of math and reading instruction on more easily tested 'basic' skills, with decreased focus on less-easily tested 'higher order' skills. It has proven particularly troublesome to define outputs as passing a *threshold* ('proficiency'). Gains made by 'bubble' students, those just below the threshold on whom attention is concentrated because they alone have a near-term potential to cross it, may come at the expense of students who are already above the threshold and perhaps also at the expense of those who are far below it.

2) Adjusting expectations of performance for the characteristics of *inputs* has proven more difficult than anticipated. With students at different risks of failure because of their varied background characteristics, accountability and incentive systems can be credible only if sanctions and rewards can be adjusted for these variations. *Defining subgroups* and measuring their performances separately is one way, but educators have not determined how to tailor expectations by subgroup. Should teachers and schools held accountable for educating more disadvantaged students be judged by the achievement growth of these students, or by an absolute standard? And if broad subgroup definitions do not capture the considerable variation in student background characteristics, do accountability systems in education create incentives for *cream-skimming*? With school choice expanding, and subgroup definitions broad, do some schools and teachers meet public expectations by the subtle selection from at-risk subgroups of those students who are least at risk?

3) *Untrustworthy statistics* undermine the credibility of accountability and incentive systems. They would do so even if measurement of outputs and inputs could be defined more precisely. Inadequate *data reliability* is one impediment: relying on a single annual test of relatively small student cohorts in schools, NCLB tolerates large confidence intervals in score reporting. But to avoid misidentifying some low performers, others may be overlooked. Because standardized test items are too few to fully represent the curriculum, *sampling corruption* results. Teachers and schools can game accountability by over-emphasizing skills needed to answer unrepresentative test questions. More explicit *gaming* can manipulate accountability data: for example, the retention of greater numbers of low-performing students in grades prior to those being tested; the exclusion from testing of those likely to score poorly, by encouraging their absence or even by suspending them for real or alleged infractions; or the opportunistic re-

assignment of students to or from subgroups (special, second language learning, or regular education) where they can aid or do the least harm to achievement of group performance targets.

These challenges - in defining outputs and inputs and in the accuracy of data themselves - surprise many education policy makers. Some conclude that the problems stem only from the inadequacy of public educators. For example, 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.³ But this expectation denies the intent and power of incentives which, if successful, *should* redirect attention and resources to those outputs that are rewarded.

The corruption of performance incentive systems stimulated by a too-heavy reliance on quantitative measurement is not peculiar to public education. It has been extensively documented in other fields by economists, business management theorists, sociologists and historians. The present paper hopes to familiarize students of performance incentives in education with this voluminous literature from other fields. It reviews evidence from medical care, job training, crime control and other human services regarding corruption similar to what is now being encountered in public education: mismeasurement of outputs (goal distortion, and threshold standards that harmfully redirect effort); mismeasurement of inputs (imprecise subgroup definitions and cream-skimming); and untrustworthy statistics (data unreliability, sampling corruption, and other forms of gaming).*

This paper also discusses how these problems limit the use of performance incentives in the private sector, and concludes by showing that performance incentives run the risk of subverting the intrinsic motivation of agents in service professions like teaching.

* The term "input" is often used in education policy discussion to refer only to school resources, such as teachers, class sizes, textbooks, etc. This definition is too limited. If the outcome, or dependent variable, is student achievement, then the inputs, or independent variables, include not only resources but also students with their varied characteristics.

That accountability for quantitatively measured outcomes distorts organizational activity does not mean that performance incentive systems should not be employed in public education. This paper describes institutions in health care, job training and welfare administration, and in the private sector, that employ performance incentive systems, notwithstanding distortions that result. In some cases, such systems are employed because institutional leaders are insufficiently aware of the perverse consequences. In other cases, leaders have weighed costs and benefits of performance incentives, concluding that the benefits predominate. In yet other cases, leaders employ performance incentives but dilute the use of quantitative output measures by relying more heavily on subjective supervisory judgment.

So the conclusion of this paper is not that quantitative accountability and performance incentive systems have no role in elementary and secondary education, but only that educational policy makers have been insufficiently aware of the costs and benefits, and so have failed to consider properly whether, and to what extent, such systems might be beneficial. This paper documents the costs, without describing the benefits of accountability and performance incentive plans, only because there is presently a gap in the apparent knowledge base of education policy makers regarding these costs. The benefits of greater accountability for public educators are generally accepted, and enthusiasm for precise measurement of educational achievement is generous. When the costs are equally accepted, and enthusiasm appropriately tempered, the nuanced implementation of performance incentive plans may become more appropriate.*

* This paper is equally concerned with the use of quantitative measures in accountability and in performance incentive plans in education. In general, accountability plans are those that use measured results to punish educators whose work is unsatisfactory. Performance incentive plans are those that use measured results to reward educators whose work is exemplary. The problems of measurement in both types of plans are similar.

Part I - Mismeasurement of Outputs

Goal Distortion

Federal, state, and district accountability systems in education, usually holding schools responsible only for raising math and reading standardized test scores, have led to narrowing of curriculums. Untested (usually more complex) aspects of math and reading have been given less emphasis. Untested subjects such as science, social studies, art, music and physical education have also been given less emphasis.* Less attention has been devoted to key educational objectives that are more difficult to measure quantitatively, like discipline, cooperation, and character.⁴

Such skills and traits are more difficult, but not impossible to measure. When the National Assessment of Educational Progress (NAEP) was first implemented on a trial basis in 1970, trained evaluators visited schools to observe and rate 13- and 17-year olds in an exercise to develop recommendations cooperatively, to assess whether these students could "apply democratic procedures on a practical level when working in a group."⁵ Such elements were dropped from NAEP because they were deemed too expensive when initial NAEP budgets were reduced.

As Chester Finn and Diane Ravitch write in *Beyond the Basics* (2007), regarding the No Child Left Behind Act and accompanying state accountability systems:

We should have seen this coming...- should have anticipated the "zero sum" problem...: more emphasis on some things would inevitably mean less attention to others.... We were wrong. We didn't see how completely standards-based reform would turn into a basic-skills frenzy...⁶

* NCLB now requires testing in science as well, but not for accountability purposes. 'Adequate Yearly Progress' and 'proficiency' are required only for math and reading.

Yet the economics and public administration literature has, for decades, been filled with warnings about just such a result.

Clarence Ridley and Herbert Simon, for example, in 1938 noted early attempts to create comparative indices for municipal activities, including a 1923 effort by the United States Chamber of Commerce to establish contests for cities based on quantified measures of accomplishment in fire protection, public health, and traffic safety.⁷ Similar comparisons were published in the 1920s by the National Committee on Municipal Standards on street cleaning, by the International Association of Chiefs of Police on crime control, and by the U.S. Children's Bureau on health and social services.⁸ Yet Ridley and Simon observed, "[t]he most serious single problem which still stands in the way of the development of satisfactory measurement techniques is the difficulty of defining the objectives of municipal services in measurable terms."⁹ Objectives, for example, like "improve health,... or develop good citizens must be stated in much more tangible and objective terms before they adapt themselves to measurement."¹⁰

Ridley and Simon noted that before attempting quantitative measurement, questions should be addressed such as: For evaluating library services, should judgments be made about the quality of books being circulated?¹¹ For a mortality index for public health, should all lives be considered equally valuable, those of the elderly, very young children, or productive workers?¹² Simon and Ridley had something to say about measuring school effectiveness as well:

The chief fault of the testing movement has consisted in its emphasis upon content in highly academic material... The fact that a particular pupil shows a marked improvement in reading or spelling may give some indication that a teacher is improving her performance... but the use to which the pupil puts that knowledge is the only significant point in determining the significance of subject tests in measuring the educational system.¹³

And

The final appraisal of the school system must be in terms of its impact upon the community through the individuals that it trains. How effective is the school system in raising the cultural level of the community?... What is the delinquency rate in the community?... Is the economic situation improving as a result of intelligent effort on the part of the people?... What is the proportion of registered voters to the eligible voting population?...

From a practical standpoint, no one is so optimistic as to believe that all these results can be directly measured, but... serious attempts will be made in the future to devise measures which will approximate these end-products as closely as possible.¹⁴

Finally, they observed that even if such measurements could be made, the most difficult challenge was weighting the multiple goals that any public service aspires to achieve. Unless weights are assigned, measurements of individual goals, no matter how precise, cannot be used for an overall evaluation of a public service.¹⁵

Not only economists, but sociologists and political scientists as well, have been familiar with goal distortion in large bureaucracies as a consequence of accountability for measurable output or process indicators. In *Social Theory and Social Structure*, originally published in 1949, Robert K. Merton observed that large organizations attempt to get consistent behavior from bureaucrats. But this leads to a "transference of [their] sentiments from the *aims* of the organization onto the particular details of the behavior required by the rules. Adherence to the rules, originally conceived as a means, becomes transformed into an end-in-itself; there occurs the familiar process of displacement of goals, whereby 'an instrumental value becomes a terminal value'."¹⁶

In his study of *Bureaucracy* (1989), James Q. Wilson wondered why public agencies did not employ "carefully designed compensation plans" that would permit public employees to benefit, financially, from good performance. "Part of the answer," he said, "is obvious. Often we do not know whether a manager or an agency has achieved the goals we want because either the

goals are vague or inconsistent, or their attainment cannot be observed, or both. Bureau chiefs in the Department of State would have to go on welfare if their pay depended on their ability to demonstrate convincingly that they had attained their bureaus' objectives."¹⁷ We could, of course, pay diplomats based on the number of meetings they held, or the number of dinners they attended, both of which may have a positive relationship to the goal of advancing the national interest, but if we did implement such a performance-based pay system, we might find that the number of dinners increased while the national interest was ignored.

In the 1990s, attempts to hold agents accountable for outcomes were most intense in health care, both in Great Britain and in the United States. In health care, measurement and accountability seemed, at first glance, to be relatively straightforward, especially in apparently easily-defined life-and-death cases. Both countries (and several American states) created 'report cards' to compare the extent to which patients of different doctors and hospitals survived open-heart surgery. Goal distortion was the result.

In 1994, the U.S. General Accounting Office examined U.S. health care report cards and summarized experts' concerns: "[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."¹⁸ A 1995 paper by a British economist, examining performance incentives in the British Health Service to hold practitioners to account for another seemingly easy-to-measure outcome, infant mortality, concluded that the incentives spurred "tunnel vision," an "emphasis by management on phenomena that can be quantified in the performance measurement scheme, at the expense of unquantified aspects of performance... There is ...clear evidence that the emphasis on the quantifiable... is distorting the nature of maternity services, to the detriment of the non-

quantifiable objectives. Indeed, a report by the House of Commons Health Committee makes just this point..."¹⁹

In 1991, economists and management experts assembled in Berkeley for a "Conference on the New Science of Organization." The most frequently-cited paper from that conference explained why 'pay for performance' plans were rare in the private sector (notwithstanding the unsupported belief of many education policy makers that such plans are ubiquitous in private industry). The co-authors, Yale and Stanford economists, observed that incentive pay "serves to direct the allocation of the agents' attention *among* their various duties... [T]he desirability of providing incentives for any one activity decreases with the difficulty of measuring performance in any other activities that make competing demands on the agent's time and attention. This result may explain a substantial part of the puzzle of why incentive clauses are so much less common than one-dimensional theories would predict."²⁰

Economists have long considered the Soviet economy as the paradigm for goal distortion; Soviet economic activity was corrupted by the incentivized re-allocation of attention. State industrial planners established targets for enterprise production, and punished managers who failed to meet them. There were targets, for example, for the number of shoes to be produced. Certainly, increasing output was an important goal of the Soviet shoe industry, but it was not the only goal. Factories responded to the incentives by using the limited supply of leather to produce a glut of small sizes that consumers couldn't use. Planners specified the number of kilometers that freight transport enterprises should cover each month. Certainly, transporters who cover more distance can deliver more goods. But when distance alone was incentivized, the haulers fulfilled their quotas by making unnecessary journeys or driving circuitous routes.²¹ Planners

* In the economics and management literature, a 'principal' is one who establishes goals and an 'agent' is one accountable to the principal for fulfilling those goals.

specified the number of meters to be drilled each quarter by geological prospectors. Certainly, geologists who drill more holes will, *ceteris paribus*, discover more oil. But when drilling became an end in itself, geologists dug useless holes but fulfilled their quotas.²²

Some Soviet incentives retarded technological progress. Electrifying the vast country was an important economic objective, but creating incentives to increase output gave electricity managers no reason to reduce inefficiency from the loss of current in transmission.²³ Quotas for other industries set in tons created incentives to avoid developing lighter materials.²⁴ "A long catalogue of examples of this kind can readily be assembled from the Soviet specialised press,"²⁵ including a cartoon that showed a gigantic nail that extended across the entire length of a nail factory; this was the most efficient way for the factory to fulfill its monthly quota, expressed in weight, for nails produced.²⁶

Sheet glass was too heavy when it was planned in tonnes, and paper too thick... When the indicator for cloth was set in linear metres, the result was cloth too narrow for subsequent operations... One commentator reports observing women unloading bricks from a lorry, smashing many of them as they worked. If they had unloaded the bricks more carefully, their performance indicators would suffer, and also the driver of the lorry would make fewer runs, thus damaging the tonne-kilometre indicator of his enterprise.²⁷

Substituting different indicators only shifted the distortion (if the sheet glass quota was in square meters, the glass produced would be too thin), and attempts to add additional indicators were unsuccessful, since they required formulae for weighting different indicators that were impossible to specify from afar. Attempting to specify all dimensions of performance in an accountability policy led to its own problems. Soviet enterprises eventually were faced with as many as 500 distinct indicators,²⁸ beyond the ability of any local manager to juggle. Throughout the Soviet Union's existence, planners struggled with the difficulty of setting measurable performance incentives for activities with complex goals. In 1983, the most prominent Western

expert of Soviet management concluded, "That so well-known and well-studied a problem still resists solution is proof enough that it is genuinely difficult to solve."²⁹

Sanctions and rewards for Soviet managers were usually based primarily on whether fixed production quotas were met, and only to a lesser extent on the degree to which the quota was exceeded. This gave managers incentives not only to distort production goals, but to hold down production to a level just above that minimally required. This widespread Soviet phenomenon was a significant restraint on national output, and came to be known as the ratchet effect, because managers feared that if their production increased much above the minimum, or if they disclosed to central planners that higher production was achievable, the planners would increase targets the following year.³⁰

In the United States, evidence of similar goal distortion in performance incentive systems has also long been apparent. Peter M. Blau discussed it in *The Dynamics of Bureaucracy* (1955) based partly on case studies of a state employment and a federal law enforcement agency.* Initially, Blau found, state employment case workers were rated by the number of interviews they conducted. This created incentives to work quickly, but not effectively. So seven new indicators of other goals were added, including the number of job referrals and actual placements, and the ratio of placements to interviews.³¹ Note, however, that these statistical indicators used by the state employment agency for performance evaluation in the 1950s were still greater in number than the single indicator of math and reading scores presently used or commonly proposed for education accountability. And the state employment agency Blau studied prohibited supervisors from basing more than 40 percent of an employee's evaluation on quantitative indicators.³² Blau considered that the accumulation of these quantitative indicators,

* Blau did not specify the federal agency, but it was apparently one charged with enforcing federal wage and hour standards, or similar rules.

when combined with supervisor evaluations, solved the perverse incentive problem. Nonetheless, some perverse incentives remained. For example, when factory workers were temporarily laid-off, subject to automatic recall, employment agency case workers improved their placement statistics by pretending to refer these workers to their previous jobs, wasting their own and the employment agency's time.³³

When, more recently, the federal and several state governments have reported on death rates by hospital and physician, they have hoped to persuade heart patients to choose more effective providers. But federal legislation has more than one goal for end-of-life care. One is certainly to prolong life, to reduce mortality. But a second goal of federal legislation is to require hospitals to provide information about living wills so that terminally ill patients can avoid, if they wish, artificial life-prolonging technology. The two goals are both important – reducing mortality, and providing the option of having a more dignified experience when death is inevitable. The two goals can be reconciled only by on-site judgment of physicians and families who weigh subtle factors. When the government rewards hospitals only for reducing their mortality rate, it undermines its other goal of encouraging the use of living wills.³⁴

Goal distortion has also infected efforts of private insurance plans that hoped to establish quantitative measures of physician quality. In California, for example, the Pacificare health plan established a Quality Improvement Program with bonuses for medical groups whose enrolled patients had high rates of screening for cervical and breast cancer, tests for blood hemoglobin and cholesterol, appropriate immunizations, appropriate prescriptions for asthma medication, and several other advisable practices. In general, medical groups improved on these measures after incentive payments were introduced. Experts expected that these improvements would spill over to improvement on other preventive measures which were not rewarded financially. But when

the attention of primary care physicians was channeled by these performance incentives, their groups' performance on important non-incentivized procedures deteriorated. Fewer patients received recommended screening for chlamydia, fewer diabetic patients received recommended eye exams, and antibiotics were prescribed less appropriately.³⁵ Concluded health economists who studied the PacifiCare and similar systems: "Inevitably... the dimensions of care that will receive the most attention will be those that are most easily measured and not necessarily those that are most valued."³⁶

A recent national survey of general internists found a majority who believed that the publication of such quality indicators "will divert physicians' attention from important but unmeasured areas of clinical care."³⁷

In 2002, following highly publicized cases of mistreatment of the elderly in nursing homes, the federal Centers for Medicaid and Medicare Services (CMS) established a report card, the Nursing Home Quality Initiative (NHQI), requiring nursing homes to report publicly whether they adhered to 15 recognized quality standards – for example, the percent of residents who have pressure sores (from being turned in bed too infrequently). These public reports were intended to, and had the effect of providing information about relative quality to consumers who were selecting nursing homes for themselves or their elderly relatives.

However, administrators of nursing homes, and nurses caring for the elderly, must balance many more than these 15 aspects of quality. For example, because nurses' time is limited, if they spend more time turning patients in bed (an NHQI) standard, they may have less time to maintain hygienic standards by washing their hands regularly (not an NHQI standard). Although the NHQI, intended for consumers, is limited to 15 standards, CMS monitors some 190 measures (such as hand washing) on a checklist when it inspects nursing homes for purposes of

certifying eligibility for Medicaid or Medicare reimbursement. Following the introduction of NHQI, performance on the 15 selected indicators improved, but adherence to the 190 standards overall declined, resulting in more citations for violations issued by CMS.³⁸

In effect, the introduction of the report card created incentives for nursing homes to re-balance their many daily tasks. When CMS selected the 15 standards for public reporting to consumers, it had to consider not only whether the selected standards were important, but also whether they were easily measurable and understandable by consumers as well as by medical professionals. Because the 190 measures for certification are unweighted, it is not possible to say whether an increase in the raw numbers of citations for violations means that overall quality decreased as a result of the government's report card project. But it may have.

There has been similar goal distortion in Great Britain, where governments of Margaret Thatcher and Tony Blair attempted to improve the quality of health care by establishing numerical targets which physicians and hospitals must meet. These included maximum waiting time targets for elective surgery, for emergency room treatment, for physician appointments, and for ambulance response. Providers did improve their performance in these regards, but unmeasured aspects of care deteriorated. Reducing maximum waiting times (to two years) for elective surgery required surgeons to spend more time on surgery and less on post-operative care, which is unmeasured in the accountability system.³⁹ Because surgical urgency is on a continuum, not neatly divided between elective and urgent procedures, the target for elective surgery caused practitioners to make relatively minor procedures (some cataract surgeries, for example) a greater priority, and more serious, but not quite urgent procedures a lesser priority; in that way all surgeries could be performed within the target time frame.⁴⁰ A consequence was that

average waiting times for surgery increased, to achieve the target that all surgeries be performed within two years.⁴¹

To compare the performance of maternity services, and encourage mothers to use those of higher quality, the British Health Service published comparative data on providers' perinatal mortality rates - the rate of infant deaths in the period immediately before and after birth. This is certainly the most easily quantifiable outcome of obstetrics. But there are other objectives as well, including reducing the severity of handicaps with which high-risk infants survive, and providing a more comfortable and competent experience for pregnant mothers. These objectives, though more difficult to quantify, require maternity services to devote more resources to prenatal care (usually community-based, not in-hospital). The incentive system, publishing only the quantifiable perinatal mortality rate, affected how the maternity services balanced their efforts between community-based prenatal care and in-hospital deliveries. Maternity services reduced resources devoted to the former at the expense of the latter, possibly resulting in worse developmental outcomes for live births – more low birthweight births, more handicapped survival, more learning difficulties and behavioral problems for children, for example.⁴² Holding maternity services accountable exclusively for live births may also have created incentives for clinics to advise termination of high risk pregnancies before the perinatal period begins, because only late-term abortions would be counted against the services in calculation of the mortality rate.⁴³

In the U.S., the Job Training Partnership Act (JTPA) of 1982 required local agencies to provide training to two distinguishable groups: those who "can benefit from" from training and those who "are most in need of training opportunities."⁴⁴ But workers most in need of job training and placement (for example, the long-term unemployed) may differ from those most

likely to benefit (the short-term unemployed). When the federal government offered financial rewards to agencies that had better records of placing workers in jobs, it created a perverse incentive to recruit and train only those most easy to place. There was no reward for training those most in need, and so this goal was downplayed by many agencies.⁴⁵

The Department of Labor measured successful performance by the employment and wage experience of trainees 90 days after the completion of formal training. This, however, created incentives for agencies to place workers in lower skilled and shorter-term jobs, provided only that these jobs lasted at least 90 days.⁴⁶ Training for long-term stable employment required more resources, and success rates were somewhat lower, although placement in such long term stable employment was also a goal of JTPA. The federal program could have reduced goal distortion by extending the monitoring program beyond 90 days, but this would have been more expensive and the Department of Labor was unwilling to devote resources to this endeavor.⁴⁷

A 1989 analysis, prepared for the Pennsylvania Council on Vocational Education by the consulting firm SRI International, concluded that JTPA performance standards had resulted in decreased services for the hard to serve.⁴⁸ 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..."⁴⁹

The performance incentive plans of both JTPA and its successor job training program, the Workforce Investment Act (WIA) of 1998, required local agencies to demonstrate continuous performance improvement each year. It created incentives for a Soviet-style ratchet effect: states establishing deliberately low initial targets for their training agencies, to ensure room for subsequent improvement. The federal government attempted to monitor this behavior, and in at

least one case (that of North Carolina) required a higher baseline standard when it seemed to be set deliberately low to make future growth easier.⁵⁰

A long-discredited performance incentive for police officers has been arrest quotas or, for traffic officers, ticket quotas. The most obvious flaw is encouragement of false arrests.⁵¹ As inevitable a result, however, is goal distortion. Although arresting those who commit crimes or misdemeanors is an important police function, also important are less-easily measured activities, such as patrols to inhibit crime or interventions to resolve conflicts. All crimes are not equally important, but arrest quotas encourage police to focus on easy but less important arrests at the expense of difficult but more important ones.⁵² Thus, criminologists typically advise against using such a performance incentive system and most sophisticated police departments have abandoned such systems.⁵³ In 1966, the criminologist Jerome Skolnick wrote, "The goals of police and the standards by which the policeman's work is to be evaluated are ambiguous... Even within the ranks of police specialists there is no clear understanding of goals," making judgment about effectiveness based on a simple quantitative indicator bound to distort police priorities.⁵⁴

Nonetheless, the appeal of simple management-by-numbers remains irresistible to some police chiefs. 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.⁵⁵

A curious example of goal distortion comes from an incentive system for bus drivers in Santiago, Chile. Most bus drivers worldwide are paid a flat wage. And almost everywhere, passengers complain of waiting too long for a bus to come, only to have several arrive together. To prevent this, Santiago pays most bus drivers on an incentive system (per passenger); the

authorities reasoned that if bus drivers were paid per passenger and found themselves too close to the previous bus, they would slow down, to give additional passengers time to congregate at bus stops. The result would be better service from more evenly spaced buses. So many (but not all) bus companies began to pay drivers per passenger carried.

The system works; the typical Santiago passenger waits 13 percent longer for a bus where drivers are paid a flat rate than for one where drivers are paid on an incentive contract. But instead of slowing down to allow passengers to congregate at a stop, many incentive drivers speed up, to pass the bus in front and thus collect passengers before another driver does so. Incentive contract drivers thus have 67 percent more accidents per mile than fixed wage drivers. Passengers complain that buses on incentive contracts lurch forward as soon as passengers board, without their having a chance to sit.⁵⁶

Bus drivers have to balance several goals – delivering passengers to their destinations, safety, and comfort. By creating a quantitative incentive only for the first, Santiago bus companies undermine the others.

As Donald Campbell observed, perhaps the most tragic example of goal distortion was the work of a former Harvard Business School professor, financial analyst, and business-executive-turned-public-official, who believed strongly in quantitative measures of success. In the Vietnam War, Secretary of Defense Robert McNamara demanded reports from his generals of relative 'body counts' of Americans and North Vietnamese. It is, of course, true that, just as high reading test scores are usually a reliable indicator of reading proficiency, relative casualties are usually a reliable indicator of the fortunes of nations at war; a strong inverse correlation between a nation's casualties and its success in the broader political and economic objectives of warfare should normally be expected. But an army can be corrupted if imposing casualties

becomes an end in itself and if the performance of local commanders is judged by their achievement on this quantitative and relatively easily measured indicator. Generals or their civilian leaders may then lose sight of the political and economic objectives. In Vietnam, as body counts became the objective itself, generals attempted to please their superiors by recording fewer deaths than those of the enemy. As it was impossible to hide American deaths from the political leadership, generals found ways to inflate the numbers of enemy deaths. In some cases, death became an end in itself, in other cases the categorization of deaths was corrupted (for example, by counting civilian as enemy deaths) or the numbers simply exaggerated. High enemy body count numbers led American leaders to believe the war was being won. These leaders confused our superiority in body counts with our achievement of political and economic objectives. The war was then lost.⁵⁷

Performance Thresholds

No Child Left Behind requires each state to establish a proficiency cut-off on its standardized tests of mathematics and reading, and implement sanctions for schools where increasing numbers of students do not score above this cut-off. This approach to accountability has created incentives for teachers to focus their instruction on students just below the proficiency point.* Referred to as 'bubble kids' (a term borrowed from poker, from college basketball, and perhaps from elsewhere as well, to refer to those at the cusp of qualification[†]), these students' small improvement can have a disproportionate impact on a school's reported success. NCLB offers no incentive for teachers to focus instruction on students whose ability level is already above the threshold, nor is there great incentive to focus instruction on students whose ability level is far below that point. When teachers respond to incentives by ignoring the latter group, they undermine the stated purpose of the policy, by withdrawing learning opportunities from the most disadvantaged children in order to concentrate scarce resources (teachers' time) on children whose improvement will be rewarded.⁵⁸

That accountability for achievement-at-the-bubble provokes such corruption should be no surprise, for it has done so in other well-documented accountability systems. When Pacificare instituted its Quality Improvement Program for medical practitioner groups, it established

* More precisely, the incentive is to focus instruction only on those students who are below but close enough to proficiency to have a realistic chance of achieving it, *and* who are members of demographic subgroups that are in danger of not making adequate yearly progress, as the law defines it. Here is how one teacher in Montgomery County, Maryland described the instructions her principal issued: "We were told to cross off the kids who would never pass. We were told to cross off the kids who, if we handed them the test tomorrow, they would pass. And then the kids who were left over, those were the kids we were supposed to focus on" (de Vise, 2007).

[†] In poker, the 'bubble' is the point where the next player out will not win any money, but all previous players have won some money. In NCAA basketball, 'bubble teams' are those just below the cut-off for making the play-offs when the tournament is a short time in the future; a small increase in performance, or the unexpected collapse of a higher ranked team, could push a bubble team over the threshold and into the play-offs.

financial rewards for groups whose achievement on several measures (such as the share of patients who were screened for cervical or breast cancer) was at or above the 75th percentile of all groups in the base year. As a result, very little bonus money went to groups that were initially below the threshold, even if they showed the greatest improvement, because few of these groups were close enough to the 75th percentile threshold to pass it, even after great improvement. Most (three-quarters) of the incentive funds went to groups that showed little improvement, but which were already above the threshold.⁵⁹ The administrators of the Quality Improvement Program adopted this approach because they feared that if bonuses went to rapidly improving groups without regard to a threshold or cut-point, they might be accused of rewarding poor performance.⁶⁰

Health care economists have observed that such commonplace threshold systems seem to be inconsistent with the stated goal of improving health care quality.⁶¹ They give providers whose performance is far below the target no incentive to improve,⁶² and give providers whose performance is far above the target a perverse incentive to overlook deteriorating performance, as long as it does not fall below the minimum threshold.⁶³ Yet if practitioners were paid instead for improvement, or gains, without regard to performance levels, providers that were already high-quality could be penalized – or at least not rewarded – because of ceiling effects.⁶⁴

Threshold systems in health, as in NCLB, have demonstrated other perverse consequences. Noted above was that Britain's National Health Service published the perinatal mortality rate as a performance indicator for maternity services. Perinatal mortality is arbitrarily defined as including stillbirths that occur after the first 28 weeks of pregnancy. But determination of the date of conception is never precise or certain. British obstetricians were able to improve their performance indicators in borderline cases by reporting that mortality occurred

before, not after the 28-week cutoff. In that case, mortalities were considered abortions, not stillbirths, and not counted in the published perinatal data.⁶⁵

The British Health Service also established a target that no patient should sit in an emergency room for more than four hours before seeing a physician. In response, hospitals dramatically improved their consistency in meeting this threshold. But average waiting times sometimes increased as a result, and health care deteriorated. Previously, the highly publicized cases that gave rise to the target were mostly patients with relatively minor injuries or illnesses who were forced to wait on the infrequent (but not unheard of) occasions when emergency rooms were overwhelmed by more serious cases. To meet the new threshold requirement, hospitals ensured that patients with less serious problems were seen before the four hours expired, but to accomplish this, patients with more serious problems had to wait somewhat longer. In some cases, hospitals diverted personnel from other care to work in the emergency room, because normal hospital care was not associated with quantifiable targets. In other cases, patients were unnecessarily transferred to regular hospital wards, to free up space in emergency rooms.⁶⁶ As a "Working Party on Performance Monitoring in the Public Services" of the Royal Statistical Society observed, the accountability target thus undermined the ethics of medical professionals, who are trained to establish priorities based on need.⁶⁷

Because the four-hour waiting standard did not begin until patients actually arrived at an emergency room, some ambulances parked and did not discharge patients to an emergency room until advised that the hospital could now see a patient within four hours. This gaming had detrimental effects on the delivery of health care, as fewer ambulances were available for dispatch to pick up seriously ill patients.⁶⁸

Yet another British threshold was that no patient should wait more than two years for elective surgery. Providers sometimes fulfilled this target by establishing a "pending list" – a waiting list to get on the waiting list – to ensure that time spent on the official waiting list would not exceed two years.⁶⁹ Another threshold standard was that patients should be able to see their primary care physicians within 48 hours of making appointments. Some physicians met this accountability threshold simply by refusing to schedule appointments more than 48 hours in advance.⁷⁰ When asked about this at a press conference, Prime Minister Tony Blair said it was "absurd" to think that doctors would do such a thing, but his Health Secretary later confirmed that this was, indeed, a perverse consequence of the accountability target.⁷¹

The National Health Service also required that ambulances must respond to 75 percent of emergency calls within eight minutes. The incentive system was successful, and many ambulance services dramatically improved their response times. Nonetheless, a data plot of response times showed a spike at eight minutes and a big drop immediately after. There was little improvement in response times that were already under eight minutes, but apparently ambulance services figured out how to get more nine minute calls to the emergency room just a little bit faster. In some cases, this was achieved by relocating ambulance depots to more urban areas; patients in rural areas had to wait much longer, but a larger number of urban patients got in just under the threshold.⁷² Perhaps this was a positive result of the incentive system, but it was not an intended one.

Observing these problems, the Royal Statistical Society working group concluded, "it is unsmart to base waiting time targets on time cut-offs..., unless data on the distribution of waiting times are also collected." Holding health care providers accountable for improvement in averages

creates different incentives from those resulting from holding providers accountable for achieving cut-off points.⁷³

The Workforce Investment Act established individual client thresholds as the basis for incentive payments. Service centers were rewarded if workers who enrolled for training found jobs whose pay was higher than the jobs they held previously. In other words, the threshold for each worker was that worker's previous pay rate. This system, however, created a disincentive to serve one group of workers who badly needed re-training and whose plight inspired the Act itself: workers who had been laid off from their previous jobs and were not 'voluntarily' seeking retraining. Often, these were manufacturing jobs, and the dislocated workers were unlikely to find new employment that paid as well as the closed factory. The problem was exacerbated because often, as a manufacturing plant prepares for layoff or closure, it schedules overtime for workers who are in the last groups to be laid-off. These workers had extraordinarily high previous earnings targets to meet. As a result, WIA service centers made little effort to recruit such workers, and instead concentrated on enrolling those who were previously unemployed or in marginal jobs, where placement in jobs only slightly better paid would count towards meeting the target.⁷⁴

Accountability for thresholds has also created opportunities for gaming in police work, illustrated by President Richard M. Nixon's 'war on crime' which included incentives to reduce serious larceny - defined as larceny having a threshold value of \$50 or more.

As a 1968 presidential candidate, Nixon made crime an issue and once in office, made reduction of crime in the District of Columbia a priority; he also focused national attention on reducing crime in other cities. Crime statistics from the Uniform Crime Reporting (UCR) system were publicly reported by the Federal Bureau of Investigation (FBI). The UCR generated a sum

of crimes in seven categories considered most important: murder, forcible rape, robbery, aggravated assault, burglary, larceny of \$50 and over, and auto theft. The District subsequently posted a significant reduction in crime, as did other cities where the UCR was publicized.⁷⁵

Mostly, the crime reductions were apparently realized by moving crime classifications across a threshold. The biggest reductions were in larcenies of \$50 and over in value. When police take a crime report, valuing the larceny is a matter of judgment. Indeed, it is a matter of practical discretion whether police turn a citizen complaint into a crime report at all.⁷⁶ Although the District's UCR index number declined, the number of larcenies valued at about \$49 went up. Larcenies which, prior to publicizing the UCR index, would have been valued at \$50 or more were now valued at less.

Other definitions of crime were also manipulated, where thresholds were more a matter of police subjective judgment. According the FBI, burglary is defined as forcible entry with the intent of committing a felony or theft. As police made more judgments that this was not the intent, the number of burglaries declined while malicious mischief or vandalism (neither of which were UCR index crimes) increased.⁷⁷ Around the country, policemen told reporters and researchers that they were under orders to downgrade the classification of crimes, in order to show progress in their cities' UCR index numbers.⁷⁸

As Donald T. Campbell summarized: "It seems to be well-documented that a well-publicized, deliberate effort at social change - Nixon's crackdown on crime - had as its main effect the corruption of crime-rate indicators, achieved through underrecording and by downgrading the crimes to less serious classifications."⁷⁹ Perhaps attention to these numbers also created incentives for better policing. It is difficult to tell. Today, police officers apparently continue to downgrade crime reports, crossing thresholds to improve their perceived efficiency.⁸⁰

Part II - Mismeasurement of Inputs

Defining Subgroups, or Risk Adjustment

Contemporary education accountability systems, at the federal and state levels, imply that measurement of school and teacher effectiveness must take some account of the different characteristics of students. Typically, that means reporting separately on performance of students of different racial or ethnic origins (white, black or Hispanic) and on performance of students from different economic circumstances (whether eligible for lunch subsidies).

But compared to other sectors – health and job training in particular – the adjustments made by educational accountability systems are both perverse and primitive. They are perverse because, whereas in other sectors, identification of subgroups serves the purpose of differentiating outcome expectations for members of different subgroups, in education the subgroups are usually defined so that educators can be held accountable for eliciting the same performance from members of different subgroups. The adjustments in education are primitive because the few subgroup definitions (usually race, Hispanic ethnicity and eligibility for lunch subsidies) are much too broad to capture relevant differences between students. Policymakers in other sectors recognize that more sophisticated background controls are required before it is possible to develop reasonable expectations for group or individual performance.

This was an important problem identified by Clarence Ridley and Herbert Simon in their 1938 study of municipal functions. To compare the effectiveness of fire departments in different cities or years, they found it impossible to use simple quantitative measures, such as the value of fire losses in a year, or the number of fires per capita. From one year or place to another, there might be a change in the amount of burnable property or in the proportion of industrial property;

a more severe winter, or "a multitude of other factors beyond the control of the administrator [that] would have an important effect upon the loss rate."⁸¹

And Ridley and Simon considered fire the easiest of municipal activities to measure.⁸² Comparisons of police effectiveness, they argued, had to account not only for racial and ethnic differences in populations, but the quality of housing, economic conditions, the availability of "wholesome recreational facilities," how the courts and penal system was administered, and "other intangible factors of civic morale."⁸³ Evaluation of public health workers' performance had to adjust for similar factors, as well as for climate, epidemics and other chance fluctuations in population health. Construction of a mortality index for measuring the adequacy of public health departments must distinguish "only those diseases which are partly or wholly preventable through public health measures."⁸⁴

Today, the process of controlling for patient differences in measuring the effectiveness of health care is termed 'risk adjustment,' connoting that doctors' or hospitals' effectiveness cannot be judged without knowing the specific risks faced by their particular patients – because of previous medical history, detailed socioeconomic and family circumstances, and present symptoms.

A 2004 RAND survey comparing accountability in education and other sectors concluded:

An important difference between health and education is that large bodies of data on health risks have been collected through screening tests, physical exams, histories, diagnostic tests, etc. and documented in inpatient and outpatient medical records. Educators, however, do not have a similar body of risk data, and they face complicated access issues if they try to collect it. Among those issues are privacy protection and the costs of accessing, abstracting, and verifying accuracy. The risk data currently available in education may be inadequate to support a proposed accountability system.⁸⁵

Yet although risk adjustment in medicine is far more sophisticated than controls for subgroups in education, health policy experts still consider that the greatest flaw in medical incentive systems is their inability to adjust performance expectations adequately for risk.

The Health Care Financing Administration (HCFA) initiated its performance incentive system for cardiac surgery in 1986, reporting on death rates of Medicare patients in 5,500 U.S. hospitals, using a multivariate statistical model to identify hospitals whose death rates after surgery were greater than expected, controlling for patient characteristics. Hospital administrators and medical professionals protested loudly, claiming that the patient characteristics incorporated in the model were not adequate to account for the challenges faced. The institution labeled as having the worst death rate even after statistical adjustment (88 percent of all Medicare patients died) turned out to be a hospice caring for terminally ill patients.⁸⁶

The following year, HCFA added a much larger number of patient-level variables. A 1990 survey of hospital administrators still found over half who claimed that the accuracy of the data as an indicator of performance was poor, and another quarter who claimed it was only fair. Only 7 percent considered it excellent or very good. The ratings were roughly similar from administrators of both hospitals that were rated as having better-than-average and worse-than-average mortality.⁸⁷ The statistics were too complex for some hospital administrators to understand; this made it difficult for administrators to infer lessons from the results for their institutions' practices, and rendered the value of these data for an incentive system questionable.⁸⁸ Other surveys of physicians in New York and Pennsylvania, states with cardiac surgery report cards similar to Medicare's, found similar results: most physicians not only believed that the incentive systems do not adequately adjust performance comparisons for patient characteristics, but stated that the failure to make accurate adjustments led them to avoid

operating on patients whose illness is more severe in ways they believed were not detected by the statistical systems. The physicians predicted that if Medicare were to implement a pay-for-performance system, avoidance of operating on high-risk patients would also result.⁸⁹

In 1991, a team of health policy researchers re-analyzed the federal Medicare data. The researchers obtained additional information on patient characteristics, enabling them to control for even more background factors than did the Medicare system. Nearly half of the 187 hospitals that Medicare had identified as being high-mortality outliers for cardiac surgery, presumably because of poor quality of care, no longer were in the high-mortality group when patient characteristics were more adequately controlled.⁹⁰ Another study compared whether these hospitals would equally be deemed outliers (with excessive mortality rates) if the statistical controls were employed using data from Medicare claims (the system used by HCFA), or using data from more detailed clinical records of the hospitals themselves. Only 64 percent of the hospitals deemed outliers with controls from the claims data were also outliers with controls from the clinical data.⁹¹

An analysis of gastro-intestinal hemorrhage cases in Great Britain found successive revisions of hospital rankings as additional adjustments for patient characteristics were applied.⁹² A study of stroke victims applied 11 alternative (and commonly used) systems for measuring severity of risk and found that the agreement between systems for identifying high- or low-quality hospitals was only fair to good; some hospitals deemed better-than-average according to some systems were deemed worse-than-average according to others.⁹³

A study of cardiac surgery outcomes in Toronto found that doubling the number of statistical adjustments for patient characteristics from 6 to 12 did not improve the prediction of relative mortality, perhaps because the additional variables were relatively evenly distributed

between hospitals.⁹⁴ Note, however, that even 6 variables in a multiple regression is from three to six times as many as anything presently attempted in education, where outcomes are typically adjusted either for race, or for free lunch eligibility, and very rarely, for both considered together.

Questions about the adequacy of risk adjustment in the HCFA report cards did not abate after the more detailed risk-adjustment methodology was applied in 1990, and although the agency had insisted that its model adequately adjusted for all critical variables, the ratings invariably resulted in higher adjusted mortality rates for low-income patients in urban hospitals than for affluent patients in suburban hospitals.⁹⁵ The Medicare performance indicator system was abandoned in 1993. Bruce Vladeck, the HCFA administrator at that time, conceded that the methodology was flawed. "I think it's overly simplistic," he told an interviewer. [I]t doesn't adequately adjust for some of the problems faced by inner-city hospitals."⁹⁶ Added Jerome Kassirer, then editor-in-chief of the *New England Journal of Medicine*, "The public has a right to know about the quality of its doctors, yet... it is irresponsible to release information that is of questionable validity, subject to alternative interpretations, or too technical for a layperson to understand..." He concluded that "no practice profile [i.e. physician report card] in use today is adequate to [the] task."⁹⁷

The General Accounting Office's 1994 study of report cards in health care observed that at least 10 identifiable patient characteristics, some involving age and socioeconomic status and some involving comorbidities or psychological functioning, might influence the outcome of health care, independent of the practitioner quality that report cards attempted to measure.⁹⁸ Although the federal Medicare report card had been abandoned while the GAO study was underway, several state health care incentive systems were still in place, as were systems adopted by private insurers. The statistical adjustments varied greatly. Pennsylvania, for example,

published a report card on hospital quality in which mortality was reported in relation to admission severity – the hospital's rating of patient risk of heart, lung, or kidney failure within the first 2 hospital days, on a scale of 0 – 4.⁹⁹ The health care industry developed several alternative statistical packages, sold to states and insurers as ways to adjust performance expectations for patient characteristics.¹⁰⁰ The best ones, however, are very costly¹⁰¹ and, as we have seen, the more sophisticated they are, the more difficulty hospital administrators have in drawing conclusions about the quality of their practice. Even so, the GAO found that no state or private insurer had been able to develop a method to adjust for patient characteristics that was "valid and reliable."¹⁰² Kaiser Permanente in Northern California, for example, published a report card that included over 100 measures of performance.¹⁰³ Yet, the GAO observed, "each performance indicator may need its own separate adjustment because patient characteristics have a unique affect on every condition and disease."¹⁰⁴ An analogy in education might be that family characteristics have a more powerful impact on reading scores than on math scores, the latter being more sensitive to school quality and the former to family intellectual environment.

It is not simply a question of identifying the proper background factors. Even if they were known, hospitals don't necessarily have all the relevant data – on a patient's previous treatment, for example.¹⁰⁵ And although research is more advanced in health than in education, risk adjustment in medicine is complicated by research uncertainty about how practice quality or patient characteristics interact to affect outcomes. For example, almost all private managed care plans use the proportion of low birth-weight babies as a measure of the quality of routine prenatal care. Yet there is no scientific consensus on the extent to which low birth-weight results from poor quality of care, or other patient characteristics over which physicians have no control.¹⁰⁶

Less-than-perfect risk adjustment has another consequence: by penalizing surgeons for operating on patients having high, but not controlled-for risk, the incentive system promotes the use of conventional strategies and the avoidance of developing new and improved, but experimental surgical techniques.¹⁰⁷ Jerome Kassirer warned that "innovation that improves clinical care must not be stifled," yet this is one of the consequences of practice conservatism that is a by-product of published medical report cards.¹⁰⁸

The Bush administration has now reinstated the Medicare accountability project, publishing the names of 41 hospitals with higher than expected death rates for heart attack patients. It plans to add pneumonia next year. Secretary of Health and Human Services Michael Leavitt acknowledges that the list still imperfectly adjusts for patient characteristics, but promises that "[i]t will get nothing but better as time goes on."¹⁰⁹ As of this writing, six states publish report cards on cardiac surgery mortality rates in their hospitals, and three publish performance reports for individual surgeons.^{*110}

For private insurers attempting to create incentives for better medical practice, distinguishing background characteristics from physician quality has also not been simple. In one case, Washington State Blue Shield had a list of routine preventive care procedures it expected doctors to perform on their regular patients. A group of physicians sued the company after it prohibited subscribers from continuing to use doctors who did not meet the insurer's standards. One physician had been banned only a year after the company gave him a \$5,000 award for being in the top 10 percent of practicing physicians in quality. It turned out he had been docked points for failing to perform a routine Pap smear on a patient who had previously undergone a

* The six states publishing hospital report cards are California, Florida, New Jersey, New York, Massachusetts, and Pennsylvania. The three publishing individual surgeon performance reports are California, New Jersey, and Pennsylvania. As of late 2006, Massachusetts was considering whether to do so.

hysterectomy. He had also been docked for not performing a routine mammogram on a patient who had previously undergone a double mastectomy. Blue Shield subsequently abandoned its performance incentive plan.¹¹¹

The inability to adjust performance expectations adequately for background characteristics has also frustrated efforts to design accountability systems in job training programs and in welfare reform. JTPA and WIA, as well as the 1996 welfare reform, Temporary Assistance to Needy Families (TANF), all seek to place clients in productive employment. The federal and state governments have tried mightily to create incentives for local welfare and job training agencies to achieve this. Yet quantitative incentive systems have floundered, as in health care, over the reality that some clients are easier to place than others.

Following the adoption of TANF, most states hired private contractors to administer at least some aspects of welfare reform. Wisconsin's program (Wisconsin Works, or W-2) was the most extensive and widely publicized, and cited as a model for those of other states. Private contractors were rewarded on the basis of participants' employment rate, average wage rate, job retention and quality (whether their employers provided health insurance), and educational activities.¹¹² However, because Wisconsin's contracts did not employ statistical adjustments for economic conditions or recipients' qualifications (for example, whether they had high school diplomas), contractors discouraged enrollment of the relatively harder-to-serve cases and profits were excessive. After each contract period, Wisconsin re-defined the incentive criteria to attempt to take account of changes in economic conditions and in contractors' opportunistic selection of clients; otherwise, meeting the state's incentive criteria became even easier (or, in some cases, more difficult). Partly, the state responded in Soviet fashion, adding five new outcome measures to its performance incentive system, but the contractors continued to make unanticipated profits

because they could continue to make judgments about potential clients that were more subtle than the state's categories could distinguish. Differences in economic conditions or participant characteristics, not quality of service, turned out to be the chief determinant of whether a private contractor received incentive payments. Eventually, Wisconsin gave up, eliminating performance standards and bonuses; this year, however, the state is re-instituting a redesigned incentive system; it is too early to say what, if any, corruption in mission may result from the new incentives.¹¹³

Analysts of TANF recently concluded that "performance measures tend to be fairly noisy signals of service value-added, in part because of the difficulty of distinguishing the contributions of service providers to participant outcomes from those of participant characteristics and external factors."¹¹⁴ As a result of these difficulties, at the federal level the Department of Health and Human Services has apparently discontinued using quantitative incentive systems to manage state TANF programs.^{* 115}

Unlike incentive systems in state TANF programs, those in both JTPA and WIA employed statistical adjustments to account for the fact that it was easier for some local agencies to place unemployed workers in jobs than for others. Agencies located in areas with booming economies, or where unemployed workers were likely to have high school diplomas, found it easier to post high placement numbers than agencies in depressed areas with large numbers of dropouts. Under JTPA, the Department of Labor itself employed a regression model to adjust for such factors in establishing a training center's goal; under WIA, the Department negotiated standards, state by state, instructing states to take into account economic conditions and participant characteristics.¹¹⁶ Nonetheless, despite such state flexibility, a General Accounting

* In Congressional testimony regarding TANF reauthorization in 2005, Assistant Secretary of Health and Human Services Wade Horn proposed that the performance incentive program be cut in half, and used only to reward employment outcomes. However, even this reduced program was not implemented.

Office survey found that most states still believed that performance levels were too high because the negotiations "did not sufficiently account for differences in economic conditions and populations served."¹¹⁷ Differences in local economic conditions, not captured by the statistical models, could be subtle – growth in new or existing businesses, for example. If the statistical adjustments had been accurate, there would be no incentive for local agencies to avoid serving difficult-to-place individuals. In the absence of accurate adjustments, however, the incentive system sabotaged the very purpose of the training acts by encouraging agencies to select only those unemployed workers who were easiest to place.¹¹⁸ The GAO concluded: "Unless the performance levels can be adjusted to truly reflect differences in economic conditions and the population served, local areas will continue to have a disincentive to serve some job seekers that could be helped."¹¹⁹

Yet with all their shortcomings, the statistical adjustments used for comparing the job placement performance of training agencies are far more sophisticated than those presently available to education policymakers. When adjusting expectations of job placement for local economic conditions, for example, JTPA and WIA use a continuous measure, the unemployment rate. Education accountability systems, however, can employ only a dichotomous measure – whether students are or are not eligible for free lunch, with family income above or below 185 percent of the national poverty rate. Yet students from very poor families have more difficult challenges than students from families at the top of the reduced-price-lunch-eligible range. In addition to having many fewer economic resources, they are also more likely to come from distressed neighborhoods, be in poor health, and have one or more parents absent. School accountability systems cannot adjust for such differences.

'Cream-Skimming'

The imprecise identification of relevant subgroups or, in other words, an inadequate risk-adjustment for the comparison of performance, creates incentives for agents to meet accountability targets by taking advantage of imperfections in the subgroup definitions or risk adjustment categories. In human services, agents do so by disproportionately selecting those clients who are easier to serve because these clients have uncontrolled-for characteristics. Whether in education, medical care, job training or other activities, agents may attempt to serve those whose risk of being unresponsive to treatment is less than predicted by the risk-adjustment or sub-group categorization system in use. This practice is commonly referred to in the economics and public administration literature as 'cream-skimming'.

In education, for example, if report cards compare schools with similar percentages of African-American students, schools have incentives to recruit African-Americans whose parents are high school graduates, because such students are easier to educate and accountability systems do not control for parent education levels. Teachers have incentives to seek assignment to schools where students are easier to educate, even if these students' superficial demographic characteristics are similar to those of students who are more difficult.* Schools of choice, though formally open to all students, use a variety of recruitment and interview techniques to discourage the enrollment of students not likely to meet accountability targets.¹²⁰ Policies to eliminate social promotion exclude more low performing students from grades being tested.¹²¹ When individual

* It has become commonplace to advocate the use of gain (or value-added) scores, rather than score levels at a single point in time, for accountability purposes, partially because it is believed that the use of value-added scores would eliminate incentives for cream-skimming in education. However, the assumption that achievement of students who begin at different points and with different demographic characteristics should, *ceteris paribus*, be expected to grow at identical rates, has never been demonstrated, and there is some reason to doubt the assumption. To identify teachers who produce greater gain scores, it is still necessary in most, if not all, value-added models to control for student demographic characteristics (Ballou, 2007; Rivkin, 2007).

teachers are accountable for test scores, internal tracking systems or disciplinary policies can result in students assigned to teachers in a non-random fashion.

These practices also have analogies in other sectors, analyzed extensively in years preceding the contemporary push for performance incentives in education.

New York State is one place where physician scorecards are issued on heart surgery mortality rates. A survey of cardiologists in that state found an overwhelming majority who were now declining to operate on patients who might benefit from the surgery, though with greater risk, in order to improve their rankings on the state reports.¹²² These cardiologists were apparently able to detect risk factors in surgery candidates that were not identified by the risk-adjustment system, and then decline to operate on patients who had these more subtle risks.

Such cream-skimming has serious consequences for the health system. Patients who arrive at a hospital suffering a heart attack have only a 40 percent chance of survival without surgery. If surgery is performed, the survival rate increases to 50 percent. But New York physicians' cream-skimming upon the advent of report cards made surgery less likely for patients whose chances of survival would have improved had the surgery been performed.¹²³

In Great Britain, too, publication of mortality data on surgical cases made surgeons more reluctant to operate on high-risk cases.¹²⁴

As a result of cardiac surgery report cards, some hospitals, more skilled at selection, got better results, while others did worse because they received a larger share of patients with the most severe illness. In New York, North Shore University Hospital in Manhasset was willing to see its 1990 ranking drop to 27 out of 30 because it was willing to operate on sicker patients referred by other hospitals.¹²⁵ In cases where time permitted, some hospitals referred patients out-of-state, where performance incentives would not operate to discourage surgery.¹²⁶ In 1989,

St. Vincent's Hospital in New York City was put on probation by the state after it placed 24th 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.¹²⁷

Pennsylvania is another state where such report cards have been published. A survey of physicians in that state also found a majority claiming that the accountability system had led to avoidance of bypass surgery or angioplasty on high risk patients.¹²⁸

Mark McClellan served as administrator of the Centers for Medicare and Medicaid Services from 2004 to 2006. A scientific report he co-authored, prior to assuming leadership of the Medicare program, concluded that 'report cards' on health care providers "may give doctors and hospitals incentives to decline to treat more difficult, severely ill patients." McClellan and his co-authors concluded that cream-skimming stimulated by cardiac surgery report cards "led to higher levels of resource use [because delaying surgery for sicker patients led to 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."¹²⁹

For hospitals overall, mortality increased, although report cards advertised that some hospitals got dramatically better outcomes. "*On net*," McClellan and his colleagues reported, "*these changes were particularly harmful...* Report cards on the performance of schools raise the same issues and therefore also need empirical evaluation" (emphasis added).¹³⁰ This paper of McClellan's was published shortly after he served the administration as a member of President George W. Bush's Council of Economic Advisers, while No Child Left Behind was being designed and implemented. Perhaps, in advising the president, Dr. McClellan weighed costs and benefits, and concluded that, "on net," the consequences of accountability in education are not welfare-reducing and are more positive than in medicine.

Cream-skimming has also been a widely-discussed characteristic of performance incentive systems in job-training agencies. In 1972-73, a professor of management conducted interviews with counselors in vocational rehabilitation programs where a reward system paid bonuses for placing disabled clients in jobs for at least 60 days. He reported that the counselors responded by competing with one another to place relatively skilled clients in jobs, and ignoring the less skilled who were harder to place.¹³¹

In addition to using statistical adjustments for risk, described above, JTPA also attempted to avoid cream-skimming by disaggregating enrollment by sub-groups. Unlike education, however, each sub-group had a unique target, considered appropriate for its unique challenges. For example, there were specific training targets for the handicapped, for racial minorities and for welfare recipients. However, these categories were too broad to defeat the cream-skimming efforts of counselors at the Private Industry Councils which administered JTPA.¹³² Counselors were able to distinguish potentially more successful trainees (for example, blacks and welfare recipients who were relatively more able than others) from within these targeted groups. Those with more education were recruited for training.

This problem is quite difficult to solve, because no matter how carefully a principal specifies targets, an agent will always know more detail than the principal about client characteristics, and will be able to cream-skim from the disaggregated pool. This is certainly true in education, where teachers know much more about their students' potential than administrators or policymakers can infer from prior test scores and a few demographic markers. "[N]ot all welfare recipients ...are identical; some are easier to serve than others, and the agent observes this information... The principal [can correct] some distortions because the agent's attention is now focused on a needier target population, but the agent will still select those applicants who

are the easiest to serve within these sub-populations..."¹³³ Literature on health care includes identical observations. Mark McClellan and his colleagues observed, "Doctors and hospitals likely have more detailed information about patients' health than the developer of a report card can, allowing them to choose to treat unobservably (to the analyst) healthier patients." If health care providers make such choices, report cards may be welfare-reducing.¹³⁴

A study of JTPA programs in Tennessee found that high school dropouts comprised 53 percent of blacks who were eligible for training (because they had income below the poverty line, were on welfare, or had certain other characteristics defining them as 'disadvantaged'). But only 23 percent of black JTPA participants were dropouts.¹³⁵ Economists who analyzed the Tennessee program concluded "even when racial and welfare targets are met, it is the most able among these groups who are chosen for help."¹³⁶ The economists considered whether this cream-skimming could be avoided by creating additional categories – for example, welfare recipients who had dropped out. The scholars concluded that creating adequate sub-category controls could not be done "short of micro-management which could destroy the prized local-responsibility and initiative features of the program."¹³⁷

James Heckman made a similar observation about the nationwide JTPA experience: There is "evidence of cream-skimming at the enrollment stage, where program staff members have the most influence. Blacks, persons with less than a high school education, persons from poorer families and those without recent employment experience are less likely to be enrolled than others, conditional on application and acceptance."¹³⁸

When JTPA was succeeded by WIA in 1998, cream-skimming of the least disadvantaged blacks and welfare recipients apparently increased, because WIA performance incentives initially weakened even the minimal demographic controls of JTPA. According to a report in

Public Finance and Management, WIA designers, apparently having failed to read the extensive economic literature on cream-skimming in JTPA, "took a problematic JTPA system and made it worse."¹³⁹ The U.S. General Accounting Office reported to Congress in February 2002 that, nationwide, "the need to meet performance levels may be the driving factor in deciding who receives WIA-funded services... Local staff are reluctant to provide WIA-funded services to job seekers who may be less likely to get and keep a job... As a result, individuals who are eligible for and may benefit from WIA-funded services may not be receiving services..."¹⁴⁰

Throughout JTPA, and now in WIA, policymakers continually adjusted the performance measures in an effort to frustrate cream-skimming behavior. As of this writing, there are 17 performance incentive standards in WIA, but each effort of the Department of Labor to solve the problem gives rise to new cream-skimming techniques.¹⁴¹ As noted above, the Soviet Union got to the point where some enterprises were judged by as many as 500 standards.

The state of Michigan, however, tackled the problem in a different way. It substantially reduced penalties and weakened its performance incentive system; the number of registered participants then increased.¹⁴²

In education, the availability of only the most gross controls (for race and lunch-eligibility) permits frequent claims that some schools (in some cases regular schools, in some cases charter schools) outperform others with 'similar' percentages of disadvantaged students. It is usually impossible to tell whether such schools truly are superior because, like doctors and welfare or job-training caseworkers, local school officials can observe and select on characteristics that are unobservable in the data used for performance reporting.¹⁴³

Part III - Untrustworthy Statistics

Data Reliability

Poor data reliability has been an impediment to the development of accountability and incentive systems in education. Sample sizes are small (for teacher accountability, classes; or for school accountability, cohort and cohort sub-group sizes) so reliance on a single test score can generate inaccurate results because of unrepresentative samples or because of random external events that influence test-taking conditions. Attempts to hold schools or teachers accountable for value-added, or score gains over time, exacerbate the reliability problems, because they compound errors in the beginning and ending test scores.

In the summer of 2001, when the Bush administration and Congress were designing the new federal *No Child Left Behind* requirements, Thomas Kane and Douglas Staiger circulated a paper showing that the proposed system would result in many of the wrong schools being rewarded or punished solely because of these statistical sampling problems.¹⁴⁴ The paper was so persuasive that the introduction of the bill was held up for several months while administration and congressional experts tried to solve the problem. They couldn't. But they introduced the bill anyway, and the result has been some remarkable anomalies: schools rewarded one year and punished the next with no underlying change in teaching effectiveness; schools rewarded under a state's system and simultaneously punished under the federal one. Some states have avoided these aberrations by applying large confidence intervals to reported test score data, but this practice has drawn the wrath of accountability proponents.

The Kane-Staiger paper also concluded that for evaluating school quality, gain (value-added) scores over time were even less reliable than score levels at a single point in time: if there is statistical noise in any single test score measure, and the signal (or true performance) of the

same student with two different teachers has a large constant (attributable to the student's own characteristics), then combining two scores in a single measure increases the ratio of noise to information. This Kane-Staiger conclusion about greater unreliability of value-added scores confirmed an analysis performed three decades earlier by Robert E. Stake, when performance contracting enjoyed temporary popularity as an education reform.¹⁴⁵

The unreliability of performance data has been documented in other sectors as well.

Apparent changes in performance over time may, to some extent, only reflect regression to the mean. Thus, an investigation of both school and hospital performance ranking systems (called league tables) in Great Britain concluded that apparently low-ranking institutions may demonstrate apparent, but not meaningful improvements over time, while apparently high-ranking institutions may demonstrate apparent, but not meaningful performance deterioration.¹⁴⁶

Donald Campbell was particularly interested in this regression-to-the-mean problem, and illustrated it by investigating a performance incentive for automobile drivers – the speeding crackdown by Connecticut police in 1956. The state drastically increased penalties for speeding, and instructed police to be less flexible in arresting drivers for speeding; all speeding convictions were to result in automatic 30-day license suspensions, even for a first offence, with indefinite suspensions after the third conviction. In the first year of the program, traffic fatalities in Connecticut fell by 12 percent, and the crackdown was considered successful. But Campbell, examining a longer term trend line for fatalities, concluded that the treatment effect was more likely trivial. The crackdown was implemented after a spike in fatalities, which would likely have declined even in the absence of a crackdown.¹⁴⁷

Poor reliability of performance measures has been one reason that incentive systems in medicine have not been more successful. The number of cardiac surgeries performed each year,

even by specialists, may be too small for reliability.¹⁴⁸ Patient samples of primary care physicians, whose practices are more varied, must be larger still. Yet some private insurance report cards require only a minimum of 30 patients for annual performance reports.¹⁴⁹

For example, in New York State's rankings of cardiac surgeons, based on risk-adjusted patient mortality, the correlation between surgeons' rankings from one year to the next was quite small ($R^2 = .049$); nearly half of the surgeons moved from above to below average, or vice-versa, in a single year.¹⁵⁰ Even the outlier designation (unacceptably high mortality) in one year was a poor predictor of such designation in a subsequent year.¹⁵¹

Reviewing such problems in both Great Britain and the United States, the working group, referenced above, of the Royal Statistical Society recommended against publishing mortality rates for individual surgeons because

there is inherent variability in all PIs [performance indicators], however well designed, which cannot be ignored. Even if a surgeon's ability is constant and the number and case mix of patients on whom she or he operates are identical this year and next, her or his actual number of operative successes need not be the same this year and next—owing to inherent variability and despite constant ability.

The public should be educated about these issues of reliability, the working group recommended, to have a more mature debate about targets as tools to improve performance.¹⁵²

Sampling Corruption

Most educational accountability and incentive proposals rely on students' standardized test scores for information on teacher and school performance. Standardized tests, however, are only samples of student achievement. Not everything a student has learned can be included in a one hour test, nor can a test be given every day. A test is a valid measure only to the extent that students' performance on the particular test items fairly represents their performance on the broader subject matter of which the test items are a sample, and also is typical of their performance on other days of the year when tests are not administered. 'Teaching to the test' corrupts this representativeness when teachers focus their instruction only on items that are expected to be on the test, rather than covering the full curriculum; or when students are drilled immediately prior to a test in a way that is inconsistent with retaining the skills for very long after the test day. When such corruption occurs, students' answers to test items are no longer a representative sample of their skills, and performance rewards or sanctions for teachers or schools based on such test scores are inappropriately granted.

In education, sample corruption is greatest when the stakes on tests are highest. In other sectors, high stakes tests lead to similar corruption.

Television stations sell advertising at rates throughout the year determined by viewership during three designated 'sweeps' months, November, February, and May. A survey company (Nielsen) sends surveys to a sample of viewers during these months, to determine what programs typical viewers are watching and their demographic characteristics. The system assumes that programming in those months is representative of programming throughout the year for which advertising is sold. Yet the stations respond to these high stakes surveys by scheduling programs during sweeps months that are more popular, or attention-grabbing, than programs scheduled

during a typical month. Some stations even award cash prizes to viewers who watch programs at times the survey is being conducted.¹⁵³

In the 1990s, the U.S. Postal Service created a performance index to rate local postmasters and help determine the size of their bonuses. Vice President Al Gore said the index was a model for all government agencies. USPS based the index on how quickly local post offices delivered mail, determined from test letters sent out each week by the accounting firm, Price Waterhouse. But an alert clerk in West Virginia spotted the Price Waterhouse bundle when it was mailed. Supervisors then notified local post offices around the state that the letters would be arriving, and postmasters hired temporary workers to make sure the test letters got to their destinations overnight. The West Virginia district's overnight delivery score rose, but there was no improvement in their overnight delivery performance generally.¹⁵⁴ The Postal Service disciplined the West Virginia postmasters; it is unknown whether this incident reflects less egregious practices of postmasters nationwide.

Several newspapers, most notably the *New York Times*, publish weekly best-seller lists. Books on the NYT best-seller list get special displays and promotions in book stores, resulting in substantial increases in sales (and authors' royalties). The best-seller list is compiled from computerized reports sent to the *Times* from a national sample of bookstores. Publishers try to identify sample book stores, and if successful, may organize bulk purchases of a book at those outlets, thereby bumping the book up to the best-seller list. The *Times* must monitor book store sales to identify such artificial purchases that corrupt the representativeness of the index. The *Times* is not always successful. "People do try to game the list," the editor in charge acknowledged.¹⁵⁵

The Environmental Protection Agency tests all vehicles to ensure that they meet pollution control standards. For diesel trucks, the test requires the engines to run for 20 minutes while emissions are monitored. In 1998, the nation's 7 largest diesel truck manufacturers settled a civil complaint with the EPA and Department of Justice for \$83 million, the largest environmental settlement ever. Government prosecutors charged that the manufacturers had figured out how to 'teach to the test' by installing computer chips in diesel engines that kept pollution control equipment turned on during the 20-minute test, but turned off during highway driving.¹⁵⁶

Another form of sampling corruption is the intensification of effort just before the cut-off point for measuring performance, resulting in a measure that does not truly reflect ongoing performance. In schools, test drill just prior to test administration is an example of this. Television broadcaster behavior during sweeps week, or the better performance of emergency room wait times during performance measurement weeks in Great Britain is another.¹⁵⁷

This was frequently a result of Soviet planning targets as well, and it led to great inefficiency. It led enterprise managers to ignore repair and maintenance needs, so production lagged in the period immediately following the cutoff. By thus falling behind, it made such 'storming' again necessary in the subsequent period.¹⁵⁸ Soviet officials strenuously condemned this practice, yet it could not be stopped, leading a prominent student of the Soviet economy to conclude that it was actively stimulated by the production target system.¹⁵⁹

Peter Blau, in his 1955 study of federal law enforcement agents, observed a similar phenomenon: measured by the cases they completed each month, they turned their attention to the easiest cases in their portfolios as the end of a monthly performance measurement period approached.¹⁶⁰

Other Gaming

There is little distinction between much other gaming behavior, and some other unintended behaviors, already discussed, in which actors subject to quantitative incentive systems might engage. Focusing attention on bubble clients or on sampled behavior, or cream-skimming in selection, could be termed forms of gaming. There are, however, even more explicit manipulations.

In education, schools may evade accountability sanctions by re-categorizing students as members of subgroups where their unmeasured characteristics will be most helpful or least harmful, such as second language learners or special education students.^{*161} Some schools suspend low-scoring students for disciplinary infractions, before testing begins,¹⁶² or simply encourage absence, or even schedule field trips (say, to the zoo) for low-scoring students on testing days.^{†163} I described above how social promotion policies may have the unintended effect of cream-skimming, as lower-scoring students are given an extra year to prepare for tests. In some cases, however, students may be retained specifically for the purpose of raising average test scores, without regard to whether retention is educationally beneficial. These practices have been documented as responses to test-based accountability policies for many years, long pre-dating the contemporary accountability movement.¹⁶⁴ In 1993, as stakes for schools on publicly

* Under No Child Left Behind, schools count students as members of multiple subgroups (such as second language learners and disabled) simultaneously, so adding an additional subgroup classification to a student's characteristics will not ordinarily help meet adequate yearly progress requirements, unless the student can also be removed from his or her original subgroup. Subgroups also must be of minimum size to be measured, so reclassifying a student to a small subgroup may help. Under some state accountability systems, test scores of a student reclassified as disabled may not be counted.

† Although there are no nationally representative data confirming the extent of such gaming, Martha Thurlow, Ph. D., director of the National Center on Education Outcomes, states, regarding research in the early 1990s on inclusion and exclusion of students from testing: "As I presented our early findings about exclusion, teachers and parents came up to me afterward to describe what had happened to them – a principal called to suggest keeping a child home on the test day so that the child would not suffer anxiety when it was unnecessary (parent of a learning-disabled student), special education teachers talking about test day being designated as the day that they went on field trips – usually to the zoo. We had person after person relate these stories to us when we presented our data on inclusion and exclusions" (Thurlow, 2007).

reported standardized tests began to increase, the MacArthur Foundation funded a survey of gaming reponses. Robert Slavin estimated that retaining low-scoring students for an additional year was resulting in score inflation for these students of about 20 percent. Referring to the necessity to exclude very severely disabled students from testing, Lauren Resnick observed, "The minute you allow exclusion, you open up a Pandora's box of manipulation designed to make the school or district look as good as possible."¹⁶⁵

Well-known gaming behavior in higher education comes from college rankings published annually by *U.S. News and World Report*. The rankings are truly a performance incentive system; many college boards of trustees consider the rankings when determining presidential compensation and 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.¹⁶⁶

The *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 how selective a college is, determined partly by the percentage of applicants who are admitted (a more selective college admits a smaller percentage of applicants). Thus, the rankings are an ideal illustration of Campbell's law, because these factors would be quite reasonable if they were not part of an incentive system. 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, if there were no stakes to their answers, would 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, giving competitive institutions poor ratings and their own institutions outstanding ones. Likewise, a college that can only accept a small proportion of applicants is likely to be of high quality for its category, because applicants are unlikely to apply

to schools for which they know they are not qualified. But once this indicator becomes the basis for accountability, colleges have an incentive to artificially boost the number of applicants who are bound to be rejected, for example by sending promotional mailings to unqualified applicants, dropping application fees, or sending applications out to high school seniors in the mail with personal information already completed. The indicator then loses its value.¹⁶⁷

Other illustrations of gaming indicators are also well-known, even if analogies to education policy were not immediately obvious to designers of NCLB.

In 1987, the U.S. Department of Transportation began requiring 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 improve. In order not to create an incentive for airlines to hurry departures in unsafe conditions, airlines were permitted to exclude flights which were delayed because of mechanical difficulties from the calculations.

The airlines responded with two forms of gamesmanship. First, more mechanical difficulties were reported when flights were late. These were probably exaggerated. Second, airlines padded their schedules. A two-hour flight might not actually arrive any earlier, but if the schedule now allotted more than two hours, the flight's on-time performance would improve. Such gamesmanship might be costly for an airline – pilots are typically paid based on scheduled, not actual, time, and their permitted hours of flying per month have a fixed cap – but the incentives were apparently strong enough to overcome these offsetting costs.¹⁶⁸ If the original schedules had been unrealistic, then the schedule changes might have been a beneficial result,

but the system did not accomplish its stated objective, which was to improve on-time performance on previously published schedules which were purported to be realistic.

Noted earlier was that, in policing, the existence of threshold crime definitions creates opportunities for corruption. Other gaming has also been described in the criminal justice system. Donald T. Campbell, for example, observed that in the 1950s, Chicago police systematically failed to record crime reports, presumably to make crime control seem more effective.¹⁶⁹

Another illustration stems from the FBI practice of documenting clearance rates, data which become the basis for many police departments' evaluation of the effectiveness of their detectives. The clearance rate is the percentage of reported crimes that are solved (result in convictions). But police can make the clearance rate increase by shrinking the denominator – failing to take a report on a crime when a citizen complains. A more serious perverse consequence of pressure to raise the clearance rate is the practice of offering suspects a reduced charge if they confess to other crimes. Sometimes, this results in a suspect confessing to other crimes he has committed, and sometimes this results in a suspect confessing to crimes he never committed – this serves the interests of both the suspect and the police detective, because the suspect gets a lesser sentence for the crime he actually committed, and the detective gets a big boost in his clearance rate. Meanwhile, those arrested who plead guilty to only the crime for which they were arrested typically get harsher penalties than those who make such deals with the police, and who may (or may not) have committed multiple crimes. Commenting on this performance incentive system, widely in use throughout the country, the criminologist Jerome Skolnick observed in 1966: "[T]he situation in which detectives are expected to demonstrate proficiency is structured so as to invite the policeman to undermine the hierarchy of penalties

found in substantive criminal law... Thus, the standard of efficiency employed in police departments may not only undermine due process of law, but also the basic standard of justice – that those equally culpable shall be given equal punishment."¹⁷⁰

Risk-adjustment in medical incentive systems has also invited data corruption. Many background characteristics used for risk adjustment must be coded by and collected from the physicians themselves who are being held accountable for risk-adjusted outcomes. Physicians have always used great discretion in their coding. As the General Accounting Office noted in its evaluation of health care report cards, many Americans have had the experience of friendly physicians who creatively code a routine office visit to qualify for insurance reimbursement. Physicians sometimes alter coding to protect patient privacy, masking diagnoses of alcoholism, HIV, or mental illness, for example.¹⁷¹ Thus it is no surprise that after incentive systems have been put in place, physicians have used their discretion to classify symptoms which patients initially present as more severe than the same symptoms would have been classified prior to the incentive system.¹⁷² For example, after New York State began reporting death rates from cardiac surgery, the share rose dramatically of cardiac patients reported by physicians to have serious risk factors prior to surgery: cardiac surgery patients reported also to suffer from chronic obstructive pulmonary disease more than doubled, and those reported to be suffering from renal failure jumped seven-fold.¹⁷³ Since the definitions of many comorbid conditions are not precise, it is unclear to what extent physicians consciously manipulated the data. Nonetheless, 41 percent of the reduction in New York's risk-adjusted mortality for cardiac bypass patients was attributable to the apparently artificial increase ('upcoding') in reported severity of patients' conditions.¹⁷⁴

As for the British Health Service's performance indicators, a House of Commons investigation found that ambulance services met the target of eight-minute response time for life-threatening emergencies, partly by starting the clock later – for example, when an ambulance was dispatched rather than when a call was made or answered.¹⁷⁵ As it was up to the ambulance services themselves to define whether a particular emergency was life-threatening, more restrictive definitions also helped meet the performance target.¹⁷⁶

Analyses of performance incentive systems in U.S. job training programs disclose similar gaming. JTPA regulations assigned performance rewards based on trainees' employment and earning status upon graduation from the program, defined as no later than 90 days after the completion of formal training. Agencies learned to vary graduation dates to maximize their measured performance. Some trainees might be 'graduated' as soon as they found employment, even if they later (but within the 90 day period) were again unemployed. For trainees not employed, agencies delayed graduation dates as long as possible. Because performance awards were calculated at the end of each fiscal year (June 30), graduation dates for clients trained in the Spring were often timed based on whether the agencies would get a bigger reward by graduating trainees in the current fiscal year or the subsequent one, in which case the graduation date could be 'banked' until July.¹⁷⁷

JTPA agencies also gamed initial enrollment dates to maximize performance rewards. Outcomes in the reward system were counted only for job-seekers actually enrolled in a training program. This gave agencies an incentive to train clients informally, waiting to formally enroll trainees until they were certain to find employment. In other cases, since outcomes were measured 90 days after the end of formal training, agencies failed to graduate and continued formally training some clients who had little hope of finding employment, long after any hope

for success had evaporated. Such gaming behavior continued under WIA, the JTPA successor program.¹⁷⁸ As the General Accounting Office observed, "[t]he lack of a uniform understanding of when registration occurs and thus who should be counted toward the measures raises questions about both the accuracy and comparability of states' performance data."¹⁷⁹

Some agencies relied less upon fiddling with enrollment, training, or graduation dates. Instead, to maximize performance records (based on employment and earnings 90 days after training was completed), these agencies provided special services to support employment, such as child care, transportation, or clothing allowances. Such services were then terminated on the 90th day. Similarly, case managers followed-up with employers, urging them to keep recent trainees on the payroll. Such follow-up most often also ended on the 90th day. These activities were surely gaming, because they were not observed in agencies prior to JTPA establishing the 90-day standard for measuring performance.¹⁸⁰

Part IV - The Private Sector

To this point, I have mostly discussed experience with quantitative accountability or incentive systems in other public services – health care, job training, policing, and welfare – and in the Soviet Union as well. But when policymakers call for such systems in public education, they most often invoke the private sector as a model. When New York City Mayor Michael Bloomberg recently announced a 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."¹⁸¹ 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."¹⁸²

When such claims are used to justify a school- and teacher-incentive system based almost exclusively on test scores, they misrepresent how the private sector motivates employees. Although performance incentive pay systems are commonplace in the private sector, for professionals they are almost never based exclusively, or even primarily on quantitative measures of performance. In fact, while the share of all workers who get performance pay in the private sector has been increasing, the share of workers who get such pay based on objective output measures has been decreasing. This may be partly attributable to occupation shifts away from occupations that lend themselves most easily to quantitative output measurement, but not entirely. There has been a decline in commissions and piece rates within the sales and production worker categories themselves. Analysis of the National Longitudinal Study of Youth (NLSY79) reveals that 26 percent of all full time private sector workers got some form of performance pay

in 2000, up from 21 percent in 1988. But the share of workers who got output-based pay (such as piece rates or commissions) declined from 9 percent to 7 percent. The increases in performance pay were for employees who got bonuses based largely on subjective supervisory evaluations (from 14 percent to 18 percent) and for employees, mostly managers in the financial sector, who received stock options based mostly on overall firm results.^{*183} The business management literature nowadays is filled with warnings about incentives that rely heavily on quantitative rather than qualitative measures.

Even for commissioned sales workers, exclusively quantitative measures are not the rule, and when they are used, have sometimes resulted in goal distortion and corruption like that common in the public sector. In extreme cases, they produce public scandals as when, in 1989, Dun and Bradstreet faced class action lawsuits and refunded hundreds of thousands of dollars because its salesmen, paid on commission, misrepresented customers' past credit-report activity to generate demand for more expensive services.¹⁸⁴ Probably the most commonly cited cautionary tale in the contemporary business literature¹⁸⁵ concerns automotive mechanics at Sears service facilities in California in the early 1990s. The mechanics, paid on commission, alienated customers by recommending unnecessary and costly repairs. State regulators attempted to bar Sears from further auto repair business, and consumers filed a class action suit against the company. Sears abandoned the commission plan; other major retailers have also abandoned or scaled-back commission sales plans for similar reasons.¹⁸⁶

* These percentages are not mutually exclusive. I.e., a total of 26 percent got some form of performance pay, but some workers may be included both in the 7 percent who got output based pay and the 18 percent who got bonuses. The output-based pay and supervisory evaluation percentages exclude workers who got tips, although such workers are included in the 26 percent total. To the extent customers tip based on a fairly fixed standard percentage, tips should be considered a form of output-based pay. To the extent consumers tip based on an evaluation of service quality, tips should be considered a bonus payment. In 1988, 4 percent of private sector workers received tip income. This declined to 2 percent in 2000.

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 designed to prevent publicly traded firms from gaming reports of their financial conditions. Yet financial data are still too complex for laypersons to interpret - that's why investors rely on sophisticated analysts, employed to discern the underlying and often non-quantifiable potential that stock prices or other easily measured characteristics might obscure.¹⁸⁷ Analysts sometimes disagree, perhaps as often as education experts who comment on the merits of particular curricula, programs, or schools. Indeed, equities markets can only exist because these indicators are not transparent – buyers and sellers have different interpretations of what firms' financial indicators mean.

Many of the distortions and corruptions of quantitative measures in the private sector parallel those in public activities. Just as physicians re-define diagnoses when their performance incentives are risk-adjusted, factory managers re-define hard-to-monitor quality control standards to meet production targets.¹⁸⁸ Executives whose compensation is based partly on corporate earnings use their discretion in accounting practices to maximize their bonuses – among the most easily manipulated are depreciation schedules for long term assets; whether shipments to or from inventories should be accelerated or delayed at the end of an accounting period; transferring other revenues or expenses from one accounting period to another; the allocation of overhead to inventories; and whether major repair activities, research and development, and even advertising expenses, should be capitalized or expensed.¹⁸⁹ In some cases, but not all, such manipulation is criminal. But before crossing that line, managers have considerable discretion.¹⁹⁰

Within any five-year period, most firms, even the most profitable, sometimes post smaller earnings in a quarterly reporting period than they did in the previous quarter. A recent calculation shows that, based on average national and industry-specific growth rates from 1963 to 2004, no more than 71 publicly traded firms should have been expected to report 20 or more consecutive quarters of earnings growth. But in fact, 811 publicly traded companies posted at least 20 consecutive quarters of growth. The authors of this analysis conclude that most of the companies reporting such growth were manipulating their accounting to smooth earnings between quarters (for example, by timing stock repurchases), because managers' performance evaluations were partly based on posting continuous earnings growth. One perverse consequence was that when these firms' string of continuous earnings growth eventually did snap, their stock prices plummeted more than should be expected, because the incentive system had caused managers to take actions that created unrealistic expectations.¹⁹¹

Similar evidence about gaming behavior in the private sector has accumulated for many years. This behavior is often similar to that observed in the command economy of the Soviet Union, discussed above. In 1952, a prominent U.S. business theorist concluded that factory managers with quotas from headquarters generally tended to push easy jobs through the line at the end of reporting periods.¹⁹² Forty years later, a review in a prominent accounting journal concluded, "The behavioral literature on management accounting and control is replete with reports of subordinates who game performance indicators, strategically manipulate information flows, and falsify information."¹⁹³ A 1998 statistical analysis found that *most* manufacturing firms have higher sales at the end of their fiscal years, and lower sales at the beginning, solely because of the compensation schemes of both managers and commissioned salespeople.¹⁹⁴ Those selling computer systems to businesses, for example, may share or hide confidential information

about future price or technology changes in order to get customers to align their purchases with commission cut-off dates.¹⁹⁵

One recent analysis finds that salespeople in the enterprise software industry have some discretion over discounts, and use this discretion to give customers discounts to influence the timing of purchases. The salespeople generally have an accelerating commission schedule, where their commission rates increase as total sales volume increases over the course of a calendar quarter. Depending on whether a salesperson does or does not require additional sales to push him up to the next commission step rate, he may use discount policy to pull in sales that would otherwise occur later than the current quarter, or to push sales that could occur sooner out to the next. Three-quarters of total sales in the firm under study (but one typical of the industry) take place *on the last day* of the calendar quarter, a phenomenon that can be attributable only to gaming of the incentive system. On average, the excessive discounts that result are equal to about 7 percent of revenue, effectively doubling the cost to employers of the commission sales system.¹⁹⁶

Such manipulation in the private sector is not restricted to white-collar employees. Supervisors have always closely monitored factory piece workers in attempts to prevent workers from falsifying their reported production. For example, in garment or machine shops where piece-work standards can yield earnings above the legislated or negotiated minimum wage, sewers or machine operators may hide completion tickets on days or weeks when production is below the minimum standard and when minimum pay is therefore guaranteed, releasing these tickets later when they can contribute to higher earnings.¹⁹⁷

By law, German employee representatives (unions) participate with management in making investment decisions. Such co-determination can have perverse effects if the same

employees are paid for performance, not a flat rate. In such cases, employee representatives typically press for investments that increase output, but not necessarily efficiency or profitability.¹⁹⁸

And as in the Soviet plan fulfillment system, and in the incentives enacted in U.S. job training programs, the ratchet effect comes into play in cases where private sector rewards are not solely for relative differences in performance, or related linearly to production, but based on meeting a standard, or quota. Workers, or managers, have incentives to hold production down to a point just above the quota, from fears that demonstrating higher production capabilities will lead supervisors to increase the minimum standard. For this reason, when minimum standards are expected of individual workers, rather than of an entire factory, workers pressure each other to hold down effort to that just barely necessary to reach the standard.¹⁹⁹

Business school graduates are made familiar with data manipulation stemming from quantitative accountability. One Harvard Business School case study analyzes how H.J. Heinz Company managers, whose bonuses were based on continuous earnings growth, maximized their compensation by manipulating the timing of shipments or billing for services, for example by paying suppliers in one fiscal year for products or services delivered in another.²⁰⁰ Another Harvard Business School case study concerns typists employed by the Lincoln Electric Company, paid for the number of their electronically monitored keystrokes, who spent lunch hours tapping the same key over and over.²⁰¹

Even where gaming is not involved, most private sector jobs, like those in the public sector, include a composite of easily measured and less-easily measured responsibilities. One reason that individual incentive pay systems are, in fact, relatively rare in the private sector is that they encourage goal distortion like that seen in the public or quasi-public (non-profit)

sectors. Even production workers, for example, are expected to maximize unmeasured aspects of the quality of their output, to care for their machinery, and attend to preventive as well as crisis maintenance. If output quality is too poor, supervisors reject the work and piece workers get no credit. Nonetheless, a piece work system creates incentives to maximize production with output whose quality is only minimally acceptable.²⁰²

As in the public sector, adding multiple measures is insufficient to minimize goal distortion. 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 an independent survey of customers who visited bank branches. 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.²⁰³

Multiple measures are also no panacea because, as Herbert Simon and his colleague warned 70 years ago, adding additional measures to an evaluation system is relatively easy; the difficult part is weighting the various measures to develop an overall performance rating. If the weighting is not explicit and well-justified, there is a likely tendency over time to increase the weights of quantitative measures, relative to qualitative ones, because the former seem superficially to be more objective. Thus, over time, the bank's measurement system came increasingly to rely on the short term branch financial results that the multiple measures system had been intended to dilute.²⁰⁴

Because of widespread gaming of purely quantitative incentive systems, most private sector systems blend quantitative and qualitative measures, with most emphasis on the latter. This is the case even in jobs where performance seems relatively easy to quantify, such as

managing small retail outlets with few standardized items. Here, too, short term performance maximization may conflict with future prospects which rely on customer satisfaction, reputation for quality, and the reputation of other outlets over which a local manager has no control. Thus, 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 cost control, but also less easily quantifiable product quality, service, cleanliness, and personnel training. Store managers are judged by the negotiated balance of these various factors.²⁰⁵ Walmart uses a similar system. The practice of supervisors and employees negotiating quantitative and qualitative performance goals as the basis for bonus pay plans is also common for professionals in the private sector.²⁰⁶

Analyses of employee performance ratings throughout the private sector find quite low correlations between supervisory ratings based on overall performance, and quantitative indicators of employee output.²⁰⁷ Certainly, supervisory evaluations of employees are less reliable than objective, quantitative indicators. Supervisory evaluations may be tainted by favoritism, bias, inflation and compression (to avoid penalizing too many employees) and even kickbacks or other forms of corruption.²⁰⁸ That labor market outcomes seem to be correlated with employees' physical attractiveness confirms that supervisory evaluations are flawed tools for objective evaluations of performance.²⁰⁹ 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."²¹⁰ Or, "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."²¹¹ Lincoln Electric overcame the efforts of its typists to

game the piece rate system by weighting subjective evaluations of supervisors equally with piece rate data in its compensation system.²¹² To avoid gaming by corporate managers, objective and subjective evaluations of performance are generally combined, because managerial performance is too complex to be measured quantitatively, and because managers have only limited control over long-term and more meaningful firm objectives.²¹³

Management of pay-for-performance plans in the private sector is labor intensive. Bain and Company, the management consulting firm, advises clients that judgment of results should always focus on long-, 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.²¹⁴

Management literature is also filled with warnings about individual pay-for-performance plans that distribute a fixed pot of bonus money among employees. When employee compensation is based on relative performance (to other employees) rather than absolute performance, employees have incentives to sabotage the work of others with whom they are competing for rewards, collude with other workers in a cooperative effort to smooth out the rewards, apply to work in groups with the least productive fellow-employees (or try to influence the selection of less competent new employees for a work group), and avoid developing innovations that enhance overall team rather than individual productivity.²¹⁵

Concludes a Harvard Business Review article, typical merit-pay plans are "inherently a zero-sum process: the more I get in my raise, the less is left for my colleagues. So the worse my workmates perform, the happier I am because I know I will look better by comparison." One

illustrative merit-pay plan caused such worker competition that a manager reported "I was spending 95% of my time on conflict resolution instead of how to serve our customers."²¹⁶

The exception to such problems is where employees have no interaction with those who are competing for merit raises – such as chief executive officers who effectively compete with their peers in other corporations for relatively better stock market performance, but who have very limited interaction with or ability to influence those peers.²¹⁷ Teachers in schools, however, are more interdependent than this for their effectiveness.

Most private (as well as public) sector jobs have outcomes which are partially attributable to individual effort, and partially attributable to group effort. For this reason, individual merit pay plans are relatively rare in the private sector; the greater the relative proportion attributable to group effort, the rarer are individual incentives. Even in manufacturing, piece rate systems are not the rule because they create incentives for workers “to shift their attention from the team activity where their individual contributions are poorly measured to the better measured and well-compensated individual activity.”²¹⁸

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. "Management by numerical goal is an attempt to manage without knowledge of what to do, and in fact is usually management by fear," Deming insisted. Only good (subjective) leadership, not restricted to mechanical and quantitative judgment, can maximize long-term results.^{*219}

* Deming was not hostile to quantitative analysis where he thought it appropriate. Deming advocated analysis of what contributes to quality and performance through statistical modeling.

A corporate accountability tool which has more recently grown 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. These generally refer only to past performance, but future performance relies not only on a track record of financial success, but 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."²²⁰ Each of these should be incorporated, and measured if possible, in an organizational accountability system. The balanced scorecard developers likened exclusive reliance on financial measures for business accountability to pilots flying a jet airplane concerned only about airspeed, or altitude, or fuel use, rather than all of these simultaneously as well as many other factors.²²¹ In the balanced scorecard approach to business accountability, quantifiable measures should be supplemented by judgments about the quality of organizational process, staff quality and morale, and customer satisfaction. Evaluation of a firm's performance should, in this theory, be "balanced between objective, easily quantifiable outcome measures and subjective, somewhat judgmental, performance drivers of the outcome measures."²²² For "best-practice firms"* employing the balanced scorecard approach, the use

of subjective judgments reflects a belief that results-based compensation may not always be the ideal scheme for rewarding managers [because] 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.²²³

* The influential work (Kaplan and Norton, 1996) describing the balanced scorecard approach relies on descriptions of illustrative firms, including Rockwater (an undersea construction company that is a division of Brown and Root, now a subsidiary of Haliburton), Analog Devices, FMC Corporation, and five pseudonymous firms in the banking, retail, petroleum and insurance industries. Other balanced scorecard case studies are included in Kaplan and Atkinson, 1998, pp. 380-441.

Curiously, the federal government has adopted a balanced scorecard approach, simultaneously with its quantitative outcome-focused Government Performance Results Act and its exclusively quantitatively-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.²²⁴ Quantitative performance indicators play only a small role in the Department's 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," "customer satisfaction and loyalty" are difficult to quantify. Other criteria, relying heavily on qualitative evaluation, comprise the other 550 points, such as "how do senior leaders set organizational vision and values," "protection of stakeholder and stockholder interests, as appropriate," etc.²²⁵

From a belief that Baldrige principles of private sector quality could be applied as well to health and education institutions, these were added in 1999. 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."²²⁶

The most recent Baldrige award in elementary and secondary education went in 2005 to the Jenks (Oklahoma) school district. In making this award, the Department of Commerce cited the 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.²²⁷ Yet the next year, the Jenks district was deemed by the federal Department of Education to be in need of improvement under the

provisions of NCLB, because Jenks' economically disadvantaged and special education students failed for two consecutive years to make 'adequate yearly progress' in reading.²²⁸

The approaches of the federal Departments of Commerce and Education are incoherent, at best.

Part V - Intrinsic Motivation

In 1971, Edward Deci, a social psychologist, published results of experiments with college students. In his laboratory, experimental and control groups were observed playing a puzzle game. During the process, members of the experimental group were offered monetary rewards for solving the puzzles; later, the monetary rewards were withdrawn and both experimental and control groups continued to play. But the experimental group's relative performance declined after the monetary rewards were withdrawn.

Professor Deci replicated his laboratory experiment with a field experiment of similar design. He divided students who wrote headlines for a student newspaper into experimental and control groups; the experimental group received, for a limited period, monetary rewards for the speed with which they completed their assignments. Again, performance of the experimental writers fell behind that of the controls after monetary rewards ended.

Apparently, Professor Deci concluded, the students were initially intrinsically motivated to succeed in the game or headline writing, but the introduction of monetary rewards reduced this intrinsic motivation.²²⁹ When they began to think of their goals as financial, they ceased caring as much about the intrinsic worth of the tasks.

Professor Deci did not examine the relevance of his findings to performance incentives for teachers or principals, but he did consider their implications for young children in school, examining the use of rewards (candy, extra recess, stars, tokens that can be exchanged for prizes) on student learning. Relying heavily on the work of educational psychologist Jerome Bruner, Deci concluded that such incentives may work well to improve classroom discipline. This is worthwhile, because it may not matter so much to a teacher what a child's reason for behaving might be, so long as the child behaves. And tokens may also improve test scores where only

recall is involved. But "if one wishes to help children learn to think creatively, to develop lasting cognitive structure, and to be intrinsically motivated to learn, [such] reinforcement programs will interfere with these goals and therefore will be inappropriate."*²³⁰

Social psychologists continue to debate such conclusions. But the Deci experiments have also spawned research by management theorists to see if public service employees are more likely to be intrinsically motivated than private sector employees, and thus, whether monetary performance incentives might do harm to non-profit public sector professions in a way that might not occur in the private for-profit sector.²³¹

In general, most management theorists conclude that public employees (including teachers) are relatively more motivated by a belief in the goals of their organizations, while private employees are relatively more motivated by financial rewards.²³² The General Social Survey (GSS), for example, finds that public sector employees are more likely to say that it is very important to them that a job be "helpful to society" and to "help others." Private sector employees are more likely to say that high pay, promotional opportunities, and job security are very important.²³³ Even in a survey of the engineering profession, engineers working for the federal government were more likely to value making socially useful contributions while private sector engineers were more likely to value high income and promotions.²³⁴ A survey of students entering management careers found that those entering the nonprofit and government sectors valued economic rewards less than those bound for the private sector.²³⁵ A survey of middle managers in public and private enterprises found that the public managers gave less emphasis to

* Widespread contemporary enthusiasm for performance incentives in education finds expression in New York City's new experiment to pay substantial cash rewards to low-income students for high test scores (see Medina, 2007). The experiment was designed and is overseen by Harvard economics professor Roland Fryer. Professor Fryer has no published work to date indicating how, or whether, Deci's conclusions were considered in the design of the experiment. One can imagine, however, a theory suggesting that, because the experiment is targeted on the lowest-performing students, there was little intrinsic motivation to destroy in the first place.

financial career goals and greater emphasis to worthwhile social or public service.²³⁶ (School principals would be typical of such middle managers.) "Failure to properly understand and utilize the motivations of public employees may lead in the short term to poor job performance and in the long term to permanent displacement of a public service ethic," concludes a review of such surveys in a public administration journal.²³⁷

The differences between intrinsic and monetary incentives among public and private employees are not without limit. Some public sector or nonprofit employees are attracted to their agencies by job security, not idealism. Some private sector employees are attracted to their firms by the challenges and opportunities for creative satisfaction. Surveys of the intrinsic motivation literature in management and economics journals cite, for example, the zeal with which computer engineers at Data General rose to the challenge of developing a technologically advanced product, with long hours and at low pay, described by Tracy Kidder in his Pulitzer Prize-winning account, *The Soul of a New Machine*.²³⁸ (The book was published in 1982, long before days when payoffs to stock options became an inspiration to computer engineers.) But Tracy Kidder fans will also recall Chris Zajac, the Massachusetts schoolteacher-subject of Kidder's subsequent (1989) book, *Among Schoolchildren*, who traveled to Puerto Rico during spring vacation at her own expense, hoping to better understand the cultural assumptions about education that her students brought with them to school. It is unlikely that Mrs. Zajac would have done a more conscientious job as schoolteacher if she were offered monetary rewards for improved student test scores. Indeed, it is possible that such rewards would have been detrimental to her performance, if she became persuaded that efforts are not worth making if unrewarded financially. Mrs. Zajac's balance of financial and intrinsic motivations is perhaps

more common among schoolteachers than is the balance of Data General engineers among business employees.

James Q. Wilson, in his study of bureaucracy, defined professionals as those "who receive some significant portion of their incentives from organized groups of fellow practitioners located outside the agency. Thus, the behavior of a professional in a bureaucracy is not wholly determined by incentives controlled by the agency."²³⁹ Although most experts investigating intrinsic motivation study managerial employees in federal and state bureaucracies, the considerations plausibly apply to teachers, many of whom enter the profession because of a belief in the mission of public education and a devotion to children, and whose loyalty is, in Wilson's terms, to the norms of the profession, not to their supervisors.

An important effort of school reform policy today is to increase the extent to which intrinsic rewards can motivate new teachers; the Teach for America program and the recruiting campaigns of many prominent charter schools (such as the KIPP academies) are illustrative.* The management literature suggests that performance incentive pay may work at cross-purposes with this effort.

Although contemporary consideration of performance incentives makes little reference to the danger of undermining teachers' intrinsic motivation, this was debated extensively a quarter-century ago when advocates of merit pay argued, as they do today, that greater extrinsic rewards could prevent teachers from withholding effort.†

* Teach for America and similar efforts initially attempted to attract the most academically talented college graduates into teaching, with recruiting drives at Ivy League and other elite colleges. These recruits were not likely to have entered teaching without the idealistic appeal of the recruitment effort. But as these programs have grown, they have expanded recruitment efforts beyond the initial elite college set. Chris Zajac, of Tracy Kidder's account, is more typical of the nation's schoolteachers: the daughter of a factory worker, she taught in the Irish working-class community where she was raised. Teach for America is now also recruiting teachers with idealism like Mrs. Zajac.

† For a summary of positions at that time, see Johnson 1986.

The intrinsic rewards of teaching should not be exaggerated. A conclusion that intrinsic motivation plays a large role in teaching does not imply that extrinsic (monetary) rewards are not also very important. John Goodlad's 1984 survey of teachers concluded that intrinsic rewards were more important in initially attracting young people to teaching, but extrinsic rewards grew in importance as motivators for remaining. But they did not apparently grow so much in importance to overtake the intrinsic considerations. A failure to realize expectations of efficacy was the first reason teachers gave for leaving the profession, with inadequate compensation in second place.²⁴⁰

Today, as discrimination against women in the professions abates and female college graduates have a greater choice of professional careers, school districts face teacher shortages because compensation levels are too low to attract a sufficient supply, intrinsic rewards notwithstanding.*

And it is possible, of course, that if the culture of public sector enterprises were transformed so that employees valued monetary rewards to a greater extent, and were less intrinsically motivated, performance would, on balance, improve. Perhaps institutional cultures are self-selecting, and public sector enterprises that re-oriented themselves around monetary incentives would attract different and more effective employees. But if the displaced intrinsic motivation is more powerful than monetary incentives in school teaching, shifting to pay-for-performance could have a net negative effect. Little research has been done to assess the likely risk or benefit of subverting teachers' intrinsic motivation with pay-for performance.[†]

* Because total compensation includes personal fulfillment as well as financial remuneration, it is likely that school districts will always be able to pay teachers somewhat less than firms pay comparably educated college graduates for less personally fulfilling work – but perhaps not as much less as districts pay today.

[†] The current issue of *Quality Counts*, the annual magazine associated with the weekly newspaper, *Education Week*, has the theme, "Tapping Into Teaching: Unlocking the Key to Student Success." The issue

Whether extrinsic rewards undermine professional norms is an ongoing subject of debate in health care, where the report cards issued by insurance companies have come increasingly to override doctors' professional judgment. A physician complains in a recent issue of *The New England Journal of Medicine* that he's "been marked down for not having an asthma plan for someone who no longer has asthma," and observes:

U.S. doctors today have less and less to say about the care of their patients. All the complex lessons they learned in medical school are being swept aside for template care. Maybe I overestimate the next generation, but I can't imagine that young, creative people who are bright and talented enough to get into medical school will put up with this nonsense for very long. They aren't becoming physicians so they can fill in checklists and be told by a phone-bank operator what they can and cannot do for patients.

The author asks,

Do we really want doctors who are motivated by wall plaques announcing their score on some "quality improvement" initiative? Will our enthusiasm for getting high grades, being declared superior to our colleagues, and earning performance bonuses overcome our profession's traditional capacity for critical thought and reliance on empirical data?²⁴¹

Without these checklists, some patients with asthma did not have the proper treatment plan. This physician's complaint cannot itself settle whether the costs and benefits of substituting extrinsic for intrinsic motivation in medicine have been properly balanced.

has two consecutive articles, the first entitled "Advancing Pay for Performance," the second entitled "Working Conditions Trump Pay." Each article makes no reference to the other (Honawar and Olson, 2008; Viadero, 2008).

Conclusion

That performance incentive plans result in goal distortion, gaming, and corruption in a wide variety of fields is not inconsistent with a conclusion that these plans nonetheless improve average performance. Several, though not all of the analyses by economists, management experts, psychologists and sociologists, upon which this paper has relied, concluded that incentive schemes improved 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 performance incentives. The Soviet Union did, after all, industrialize from a feudal society in record time.

The survey, reported above, showing that physicians believe performance pay plans and the shaming publication of physician outcomes would result in avoidance of difficult cases and overlooking important but unmeasured aspects of treatment, also found that $\frac{3}{4}$ of physicians continued to believe that pay-for-performance is beneficial overall.²⁴² Performance incentive plans in medicine, both in the United States and Great Britain did improve average outcomes in many respects, including cardiac surgery survival rates, the most frequently analyzed procedure.²⁴³ Accountability for waiting times for elective surgery in Great Britain did reduce average waiting times, notwithstanding some other perverse consequences.²⁴⁴ One careful analysis of emergency room waiting times in Great Britain was unable to find evidence of any of the perverse consequences expected from a narrow quantitative incentive. It could be, the authors conclude, that "it is better to manage an organization using imperfect measures than using none at all."²⁴⁵ And the General Accounting Office, in its report condemning the perverse incentives resulting from report cards in health care, nonetheless concluded, "We support the report card concept and encourage continued development in the field."²⁴⁶

In education (and notwithstanding an effort of Stecher and Kirby [2004]), most policy makers who promote performance incentives and accountability seem mostly oblivious to the extensive literature in economics and management theory, documenting the inevitable corruption of quantitative indicators and the perverse consequences of performance incentives which rely on such indicators. If ignorant of this literature, proponents of performance incentives in education are unable to engage in careful deliberation about whether, in particular cases, the benefits are worth the price.

A National Academies (National Research Council) panel on “Incentives and Test-Based Accountability in Public Education” is now developing a report to make education experts more familiar with the experience of performance incentives in other fields.²⁴⁷ Several scholars cited in this report are members of the panel. Yet National Academy panels are too often ignored when the consensus of scientific judgment they bring to bear on a topic is at odds with conventional assumptions of education experts. The National Academy’s warning that consequences should never attach to a single test is an example that comes immediately to mind.²⁴⁸

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 math and reading? Will the gain in teacher quality from a performance incentive system be sufficient to justify the loss to the profession of intrinsic motivation as a driving force? How much misidentification of high or low performing teachers or schools is tolerable in order to improve the average performance of teachers or schools? How much curricular corruption, 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”?²⁴⁹

These are difficult questions that proponents of performance incentives in education must answer. As yet, the questions have been mostly unasked.

Postscript

Colleagues who read an earlier draft of this paper observed that the questions posed in the penultimate paragraph of the conclusion beg for my own answer. They observe that while the introduction and conclusion of this paper argue for balancing the costs and benefits of accountability and incentive systems that rely excessively upon quantitative measures of performance, the paper catalogues the costs, with little attention to the benefits. I agree.

However, this paper is not the place to address such issues. I incorporate the themes of this paper in forthcoming work in which colleagues and I describe an alternative (and we hope, superior) accountability system for public education. This proposed accountability system is not hostile to test scores and other quantitative measures of performance, but complements them with qualitative judgment. It includes elements now found in systems of accreditation, a minimal form of accountability.

Bibliography

- Adams, Scott J., and John S. Heywood. 2007. "Performance Pay in the US: Concepts, Measurement and Trends." 2nd Draft, November 19. Economic Policy Institute.
- Allington, Richard L., and Anne McGill-Franzen. 1992. "Unintended Effects of Educational Reform in New York." *Educational Policy* 6 (4), December: 397-414.
- Altman, Lawrence K. 1990. "Heart-Surgery Death Rates Decline in New York." *The New York Times*, December 5.
- Anderson, Kathryn H., et al. (Richard V. Burkhauser, Jennie E. Raymond, and Clifford S. Russell). 1991. "Mixed Signals in the Job Training Partnership Act." *Growth & Change* 22 (3), Summer: 32 – 48.
- Associated Press. 1993. "Rating of Hospitals Is Delayed On Ground of Flaws in Data." *The New York Times*, June 23.
- Baker, George P. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100 (3), June: 598-614.
- Baker, George. 2002. "Distortion and Risk in Optimal Performance Contracts." *The Journal of Human Resources* 37 (4), Autumn: 728-751.
- Baker, George. 2007. "Performance Indicators, Distortion, and Campbell's Law of Good Intentions." (Powerpoint). Prepared for the Eric M. Mindich Conference on Experimental Social Science: Biases from Behavioral Responses to Measurement: Perspectives from Theoretical Economics, Health Care, Education, and Social Services. Cambridge, Massachusetts, May 4.
- Baker, George, Robert Gibbons, Kevin J. Murphy. 1994. "Subjective Performance Measures in Optimal Incentive Contracts." *The Quarterly Journal of Economics* 109 (4), November: 1125-1156.
- Ballou, Dale. 2007. "Value-Added Assessment: Controlling for Context with Misspecified Models." Paper presented at the May 2, 2005 Urban Institute conference, "Learning from Longitudinal Data in Education," (revised July).
- Ballou, Dale, and Matthew G. Springer. 2007. "Achievement Trade-Offs and No Child Left Behind." Prepared for panel, "The Intended or Unintended Consequences of NCLB Accountability, Autonomy, and Choice Mechanisms on Student Academic Achievement." Annual Meeting of the American Education Research Association, April 7.

- Barnow, Burt S., and Jeffrey A. Smith. 2004. "Performance Management of U.S. Job Training Programs: Lessons from the Job Training Partnership Act." *Public Finance & Management*. 4 (3), September: 247-287.
- Berwick, Donald M., and David L. Wald. 1990. "Hospital Leaders' Opinions of the HCFA Mortality Data." *Journal of the American Medical Association* 263: 247-249.
- Bevan, Gwyn, and Christopher Hood. 2006. "What's Measured is What Matters: Targets and Gaming in the English Public Health Care System." *Public Administration* 84 (3): 517-538
- Bird, Sheila M., et al. (Sir David Cox, Vern T. Farewell, Harvey Goldstein, Tim Holt, and Peter C. Smith). 2005. "Performance Indicators: Good, Bad, and Ugly." *Journal of the Royal Statistical Society, Series A* 168 (1): 1-27.
- Birnberg, Jacob G., Lawrence Turopolec, and S. Mark Young. 1983. "The Organizational Context of Accounting." *Accounting, Organizations, and Society* 8 (2/3): 111-129.
- Blalock, Ann, and Burt Barnow. 2001. "Is the New Obsession with 'Performance Management' Masking the Truth about Social Programs?" In Dall W. Forsythe, ed. *Quicker; Better; Cheaper: Managing Performance in American Government*. Albany, NY: Rockefeller Institute Press.
- Blau, Peter Michael. 1955 (rev 1963). *The Dynamics of Bureaucracy; A Study of Interpersonal Relations in Two Government Agencies*. Chicago: University of Chicago Press.
- Bloomberg, Michael. 2007. "Mayor's Press Release." No. 375. October 17. www.nyc.gov.
- BNQP (Baldrige National Quality Program). 2007a. "Criteria for Performance Excellence." http://www.quality.nist.gov/PDF_files/2007_Business_Nonprofit_Criteria.pdf. Washington, D.C.: Baldrige National Quality Program, National Institute of Standards and Technology, Technology Administration, U.S. Department of Commerce
- BNQP (Baldrige National Quality Program). 2007b. "Education Criteria for Performance Excellence." http://www.quality.nist.gov/PDF_files/2007_Education_Criteria.pdf. Washington, D.C.: Baldrige National Quality Program, National Institute of Standards and Technology, Technology Administration, U.S. Department of Commerce.
- BNQP (Baldrige National Quality Program). 2007c. "2005 Award Winner." http://www.quality.nist.gov/PDF_files/Jenks_Public_Schools_Profile.pdf Washington, D.C.: Baldrige National Quality Program, National Institute of Standards and Technology, Technology Administration, U.S. Department of Commerce.
- Bommer, William H., et al. (Jonathan L. Johnson, Gregory A. Rich, Philip M. Podsakoff, and Scott B. McKenzie). 1995. "On the Interchangeability of Objective and Subjective

- Measures of Employee Performance: a Meta-Analysis." *Personnel Psychology* 48 (3): 587-605.
- Booher-Jennings, Jennifer. 2006. "Rationing Education in an Era of Accountability." *Phi Delta Kappan* 87 (10), June: 756-761.
- Booher-Jennings, Jennifer. forthcoming. "The Distributional Consequences of Individual and Organizational Responses to Performance Measurement Systems." Department of Sociology, Columbia University.
- Butterfield, Fox. 1998. "As Crime Falls, Pressure Rises to Alter Data." *The New York Times*, August 3.
- Campbell, Donald T. 1969. "Reforms as Experiments." *American Psychologist* 24 (4), April: 409-29.
- Campbell, Donald T. 1979. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2: 67-90.
- Campbell, Donald T., and H. Laurence Ross. 1968. "The Connecticut Crackdown on Speeding." *Law and Society Review* 3, August: 33-54.
- Carnoy, Martin, et al. (Rebecca Jacobsen, Lawrence Mishel, and Richard Rothstein). 2005. *The Charter School Dust-Up. Examining the Evidence on Enrollment and Achievement*. New York: Teachers College Press.
- Casalino, Lawrence P., et al. (G. Caleb Alexander, Lei Jin and R. Tamara Konetzka). 2007. "General Internists' Views on Pay-for-Performance and Public Reporting of Quality Scores: A National Survey." *Health Affairs* 26 (2), March/April: 492 – 99.
- Courty, Pascal and Gerald Marschke. 1997. "Measuring Government Performance: Lessons from a Federal Job-Training Program." *American Economic Review* 87 (2): 383-388.
- Courty, Pascal, Carolyn Heinrich, and Gerald Marschke. 2005. "Setting the Standard in Performance Measurement Systems." *International Public Management Journal* 8 (3): 321-347.
- Crewson, Philip E. 1997. "Public-Service Motivation: Building Empirical Evidence of Incidence and Effect." *Journal of Public Administration Research and Theory* 7 (4), October: 499-518.
- Cullen, Julie Berry, Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." NBER Working Paper W12286, June.
- Darley, John. 1991. "Setting Standards Seeks Control, Risks Distortions." *Institute of Governmental Studies Public Affairs Report*, 32 (4). Berkeley: University of California.

- Deci, Edward L. 1971. "Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology* 18 (1), April: 105-115.
- Deci, Edward L. 1975. *Intrinsic Motivation*. New York: Plenum Press.
- Deming, W. Edwards. 1986. *Out of the Crisis*. Cambridge: Massachusetts Institute of Technology, Center for Advanced Engineering Study.
- Deming Institute (The W. Edwards Deming Institute). 2007. "Teachings." <http://www.deming.org/theman/teachings02.html>. (Accessed December 1, 2007.)
- de Vise, Daniel. 2007. "A Concentrated Approach to Exams; Rockville School's Efforts Raise Questions of Test-Prep Ethics." *The Washington Post*, March 4: C01
- Dorph, R., et al. (D. Goldstein, S. Lee, K. Lepori, S. Schneider, and S. Venkatesan). 2007. *The Status of Science Education in the Bay Area: Research Brief*. Lawrence Hall of Science, University of California, Berkeley; California. http://www.lawrencehallofscience.org/rea/bayareastudy/pdf/final_to_print_research_brief.pdf
- Dranove, David, et al. (Daniel Kessler, Mark McClellan and Mark Satterthwaite). 2003. "Is More Information Better? The Effects Of 'Report Cards' On Health Care Providers," *Journal of Political Economy* 111(3), June: 555-588.
- EPA (Environmental Protection Agency). 1998. "DOJ, EPA Announce One Billion Dollar Settlement With Diesel Engine Industry for Clean Air Violations." October 22. <http://yosemite.epa.gov/opa/admpress.nsf/b1ab9f485b098972852562e7004dc686/93e9e651adeed6b7852566a60069ad2e?OpenDocument>
- Epperson, Shaun. 2007. "Jenks School Misses NCLB Standard." *The Tulsa World*, November 14.
- Epstein, Arnold. 1995. "Performance Reports on Quality – Prototypes, Problems, and Prospects." *New England Journal of Medicine* 333 (1), July 6: 57-61.
- Farhi, Paul. 1996. "Television 'Sweeps' Stakes." *The Washington Post*, November 17.
- Figlio, David, 2005. "Testing, Crime and Punishment." NBER Working Paper W11194, March.
- Figlio, David N., and Lawrence S. Getzler. 2002. "Accountability , Ability and Disability: Gaming the System." NBER Working Paper W9307, November.
- Finder, Alan. 2007. "College Ratings Race Roars On Despite Concerns." *The New York Times*, August 17.

- Finn, Chester E., Jr., and Diane Ravitch, eds. 2007. *Beyond the Basics: Achieving a Liberal Education for All Children*. Thomas B. Fordham Institute.
- Gibbons, Robert. 1998. "Incentives in Organizations." *The Journal of Economic Perspectives* 12 (4), Autumn: 115-132.
- Gibbons, Robert, and Kevin J. Murphy. 1990. "Relative Performance Evaluation for Chief Executive Officers." *Industrial and Labor Relations Review* 43 (suppl.), February: 30s-51s.
- Goddard, Maria, Russell Mannion, and Peter C. Smith. 2000. "The Performance Framework: Taking Account of Economic Behaviour." In P.C. Smith, ed. *Reforming Markets in Health Care*, pp. 138-161. Buckingham: Open University Press.
- Goldstein, Harvey, and David J. Spiegelhalter. 1996. "League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance (with discussion)." *Journal of the Royal Statistical Society, Series A* 159: 385-443.
- Goodlad, John I. 1984 (2004 edition). *A Place Called School*. McGraw Hill
- Gootman, Elissa. 2007. Teachers Agree to Bonus Pay Tied to Scores." *The New York Times*, October 18.
- Gormley, William T., Jr., and David L. Weimer. 1999. *Organizational Report Cards*. Cambridge, MA: Harvard University Press.
- Green, Jesse, Leigh J. Passman, and Neil Wintfeld. 1991. "Analyzing Hospital Mortality: The Consequences of Diversity in Patient Mix." *Journal of the American Medical Association. (JAMA)* 265: 1849-1853, p. 1853.
- Green, Jesse, and Neil Wintfeld. 1995. "Report Cards on Cardiac Surgeons: Assessing New York State's Approach." *The New England Journal of Medicine* 332 (18), May 4: 1229-1233.
- Hamermesh, Daniel S., and Jeff E. Biddle. 1994. "Beauty and the Labor Market." *American Economic Review* 84 (5), December: 1174-1194.
- Hamilton, Laura S., et al. (Brian M. Stecher, Georges Vernez, and Ron Zimmer). 2007. "Passing or Failing? A Midterm Report Card for 'No Child Left Behind'." *RAND Review*, Fall. <http://rand.org/publications/randreview/issues/fall2007/passing1.html>
- Haney, Walt, and Anastasia Raczek. 1994. "Surmounting Outcomes Accountability in Education." Paper Prepared for the U.S. Congress Office of Technology Assessment. February.
- Harris, Gardiner. 2007. "Report Rates Hospitals on their Heart Treatment." *The New York Times*, June 22.

- Harvard Business School. 1984. "H.J. Heinz Company: The Administration of Policy (A)." Case Study 9-382-034, April 1.
- Healy, Paul M. 1985. "The Effect of Bonus Schemes on Accounting Decisions." *Journal of Accounting and Economics* 7: 85-107.
- Heckman, James J., Carolyn Heinrich, and Jeffrey Smith. 2002. "The Performance of Performance Standards." *The Journal of Human Resources* 37 (4), Autumn: 778-811.
- Heinrich, Carolyn J. 2004. "Improving Public-Sector Performance Management: One Step Forward, Two Steps Back?" *Public Finance and Management* 4 (3): 317-351.
- Heinrich, Carolyn J. and Youseok Choi. forthcoming. "Performance-based Contracting in Social Welfare Programs." *The American Review of Public Administration*. Draft, October 2006.
- Heinrich, Carolyn J., and Gerald Marschke. 2007. "Dynamics in Performance Measurement System Design and Implementation." July. Draft.
- Heubert, Jay P. and Rober M. Hauser, ed. 1999. *High Stakes. Testing for Tracking, Promotion, and Graduation*. Washington, D.C.: National Academy Press.
- Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, & Organization* 7 (Special Issue): 24-52.
- Honawar, Vaishali, and Lynn Olson. 2008. "Advancing Pay for Performance." *Quality Counts, Education Week* 27 (18), January 10: 26-31.
- Horn, Wade F. 2005. "Welfare Reform Reauthorization Proposals." Testimony before the Subcommittee on Human Resources, House Ways and Means Committee, February 10. http://www.acf.dhhs.gov/programs/olab/legislative/testimony/2005/welfare_reform_testimony.html
- Hoyt, Clark. 2007. "Books for the Ages, if Not for the Best-Seller List." *The New York Times*, October 21.
- Iezzoni, Lisa I., 1991. "'Black Box' Medical Information Systems: A Technology Needing Assessment" (editorial). *Journal of the American Medical Association* 265 (22): 3006-3007.
- Iezzoni, Lisa I. 1994. "Risk and Outcomes." In Lisa I. Iezzoni, ed. *Risk Adjustment for Measuring Health Care Outcomes*. Ann Arbor: Health Administration Press.

- Iezzoni, Lisa I. et al. (Michael Shwartz, Arlene S. Ash, John S. Hughes, Jennifer Daley, and Yevgenia D. Mackiernan). 1995. "Using Severity-Adjusted Stroke Mortality Rates to Judge Hospitals." *International Journal for Quality in Health Care* 7 (2): 81-94.
- Ittner, Christopher D., and David F. Larcker. 2003. "Coming Up Short on Nonfinancial Performance Measurement." *Harvard Business Review*, November.
- Ittner, Christopher D., David F. Larcker, and Marshall W. Meyer. 1997. "Performance, Compensation, and the Balanced Scorecard." The Wharton School, The University of Pennsylvania. November 1. <http://knowledge.wharton.upenn.edu/papers/405.pdf>
- Jackman, Tom. 2004. "Falls Church Police Must Meet Quota For Tickets." *The Washington Post*, August 8: C01.
- Jacob, Brian A., 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89 (5-6), June: 761-796.
- Jaschik, Scott. 2007. "Should U.S. News Make Presidents Rich?" *Inside Higher Ed* (Insidehighered.com), March 19.
<http://www.insidehighered.com/news/2007/03/19/usnews>
- Jaworski, Bernard J., and S. Mark Young. 1992. "Dysfunctional Behavior and Management Control: An Empirical Study of Marketing Managers." *Accounting, Organizations, and Society* 17 (1): 17-35.
- Johnson, Ryan M., David H. Reiley, and Juan Carlos Munoz. 2006. "'The War for the Fare.' How Driver Compensation Affects Bus System Performance." Draft, August.
<http://www.u.arizona.edu/~dreiley/papers/WarForTheFare.pdf>
- Johnson, Susan Moore. 1986. "Incentives for Teachers: What Motivates, What Matters." *Educational Administration Quarterly* 22 (3), Summer: 54-79.
- Kane, Thomas J. and Douglas O. Staiger, 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *The Journal of Economic Perspectives* 16 (4), Fall.
- Kantrowitz, Barbara, and Karen Springen. 1997. "Why Johnny Stayed Home." *Newsweek*, October 6: p. 60.
- Kaplan, Robert S., and Anthony A. Atkinson. 1998. *Advanced Management Accounting*. Third Edition. Englewood Cliffs, NJ: Prentice Hall.
- Kaplan, Robert S., and David P. Norton. 1996. *The Balanced Scorecard. Translating Strategy into Action*. Boston: Harvard Business School Press.

- Kassirer, Jerome P. 1994. "The Use and Abuse of Practice Profiles." *The New England Journal of Medicine* 330 (9), March 3: 634-636.
- Kelman, Steven, and John N. Friedman. 2007. "Performance Improvement and Performance Dysfunction: An Empirical Examination of Impacts of the Emergency Room Wait-Time Target in the English National Health Service." *Kennedy School of Government Faculty Research Working Paper Series*, RWP07-034. August.
- Kerr, Steven. 1975. "On the Folly of Rewarding A While Hoping for B." *Academy of Management Journal* 18 (4), December: 769-783.
- Kidder, Tracy. 1989. *Among Schoolchildren*. Boston: Houghton Mifflin.
- Koretz, Daniel. 2007. "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: Biases from Behavioral Responses to Measurement: Perspectives from Theoretical Economics, Health Care, Education, and Social Services. Cambridge, Massachusetts, May 4.
- Krakauer, Henry et al. (R. Clifton Bailey, Kimberley J. Skellan, John D. Stewart, Arthur J. Hartz, Evelyn M. Kuhn, and Alfred A. Rimm). 1992. "Evaluation of the HCFA Model for the Analysis of Mortality Following Hospitalization." *Health Services Research* 27: 317-335.
- Kreps, David M. 1997. "Intrinsic Motivation and Extrinsic Incentives." *The American Economic Review* 87 (2), May: 359-364.
- Larkin, Ian. 2007. "The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales." April 12, Draft.
- Litwack, John M. 1993. "Coordination, Incentives, and the Ratchet Effect." *The RAND Journal of Economics*, Vol. 24, No. 2., Summer: 271-285.
- Lu, Susan Feng. 2007. "Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes." November 15.
<http://www.kellogg.northwestern.edu/faculty/lu/multitasking.pdf>
- McAllister, Bill. 1998. "A 'Special' Delivery in West Virginia." *The Washington Post*, January 10.
- McKee, Martin. 1996. "Discussion of the Paper by Goldstein and Spiegelhalter." In Harvey Goldstein and David J. Spiegelhalter. "League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance (with discussion)." *Journal of the Royal Statistical Society*, Series A 159: 385-443.

- McKee, Martin, and Duncan Hunter. 1994. "What Can Comparisons of Hospital Death Rates Tell Us About the Quality of Care?" In T. Delamothe, ed., *Outcomes into Clinical Practice*. London: British Medical Journal Press, 108-115.
- McMurrer, Jennifer. 2007. *Choices, Changes, and Challenges. Curriculum and Instruction in the NCLB Era*. July. Washington, D.C.: Center on Education Policy. <http://www.cep-dc.org/index.cfm?fuseaction=document.showDocumentByID&nodeID=1&DocumentID=212>
- Medina, Jennifer. 2007. "His Charge: Find a Key to Students' Success." *The New York Times*, June 21.
- Merton, Robert K. 1957 (1968 Enlarged Edition). *Social Theory and Social Structure*. New York: Free Press.
- Moore, Solomon. 2007. "In California, Deputies Held Competition on Arrests." *The New York Times*, October 5: A16.
- Morrissey, W. R. 1972. "Nixon Anti-Crime Plan Undermines Crime Statistics." *Justice Magazine* 5/6, June/July: 8 – 14.
- Mullen, P. M. 1985. "Performance Indicators – Is Anything New?" *Hospital and Health Services Review*, July: 165-167.
- Mullen, Kathleen J., Meredith B. Rosenthal, and Richard G. Frank. 2007. "Can You Get What You Pay For? Pay for Performance and the Quality of Health Care." (Powerpoint). Prepared for the Eric M. Mindich Conference on Experimental Social Science: Biases from Behavioral Responses to Measurement: Perspectives from Theoretical Economics, Health Care, Education, and Social Services. Cambridge, Massachusetts, May 4. Presentation by Meredith B. Rosenthal.
- Murray, Michael. 2005. "Why Arrest Quotas are Wrong." *PBA Magazine*, Spring. <http://www.nycpba.org/publications/mag-05-spring/murray.html>
- Myers, James N., Linda A. Myers, and Douglas J. Skinner. 2007. "Earnings Momentum and Earnings Management." *Journal of Accounting, Auditing and Finance*, 22 (2), Spring: 249-284.
- NAEP (National Assessment of Educational Progress). 1970. *National Assessment of Educational Progress, A Project of the Education Commission of the States. Report 2 – Citizenship: National Results*. November. Washington, D.C.: U.S. Government Printing Office.
- Nakashima, Ellen. 2007. "Doctors Rated but Can't Get a Second Opinion; Inaccurate Data About Physicians' Performance Can Harm Reputations." *The Washington Post*, July 25: A01.

- National Academies. 2008. "Project Title: Incentives and Test-Based Accountability in Public Education." <http://www8.nationalacademies.org/cp/projectview.aspx?key=48743>
- Neal, Derek, and Diane Whitmore Schanzenbach. 2007. "Left Behind by Design. Proficiency Counts and Test-Based Accountability." NBER Working Paper 13293. August.
- Nove, A. 1958. "The Problem of Success Indicators in Soviet Industry." *Economica* 25 (97): 1-13.
- Nove, A. 1964. "Economic Irrationality and Irrational Statistics." Chapter 16 in A. Nove, ed., *Economic Rationality and Soviet Politics; or, Was Stalin Really Necessary?* London.
- Oyer, Paul. 1998. "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality." *The Quarterly Journal of Economics* 113 (1), February: 149-185.
- Patterson, Gregory. 1992. "Distressed Shoppers, Disaffected Workers Prompt Stores to Alter Sales Commission." *Wall Street Journal*, June 1.
- Pearce, Jone L., William B. Stevenson, and James L. Perry. 1985. "Managerial Compensation Based on Organizational Performance: A Time Series Analysis of the Effects of Merit Pay." *Academy of Management Journal* 28 (2), June: 261-278.
- Perry, James L., and Lyman W. Porter. 1982. "Factors Affecting the Context for Motivation in Public Organizations." *The Academy of Management Review* 7 (1), January: 89-98.
- Perry, James L., and Lois Recascino Wise. 1990. "The Motivational Bases of Public Service." *Public Administration Review* 50 (3), May – June: 367-373.
- Pfeffer, Jeffrey. 1998. "Six Dangerous Myths about Pay," *Harvard Business Review* 76(3): 108 – 119.
- Rainey, Hal G. 1982. "Reward Preferences among Public and Private Managers: In Search of the Service Ethic." *The American Review of Public Administration* 16 (4), December: 288-302.
- Ridgway, V. F. 1956. "Dysfunctional Consequences of Performance Measurements." *Administrative Science Quarterly* 1 (2), September: 240-247.
- Ridley, Clarence E., and Herbert A. Simon. 1938, 1943. *Measuring Municipal Activities, a Survey of Suggested Criteria for Appraising Administration*. Chicago: The International City Managers' Association.
- Rivkin, Steven G. 2007. "Value-Added Analysis and Education Policy." *Brief 1*. National Center for Analysis of Longitudinal Data in Educational Research. November. http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf

- Roberts, Johnnie L. 1989. "Credit Squeeze: Dun & Bradstreet Faces Flap Over How It Sells Reports on Businesses." *The Wall Street Journal*, March 2.
- Roper, William L., et al. (William Winkenwerder, Glenn M. Hackbarth, and Henry Krakauer). 1988. "Effectiveness in Health Care: An Initiative to Evaluate and Improve Medical Practice." *New England Journal of Medicine* 319 (18): 1197-1202.
- Rosenthal, Meredith B., et al. (Rushika Fernandopulle, HyunSook Ryu Song, Bruce Landon). 2004. "Paying For Quality: Providers' Incentives For Quality Improvement." *Health Affairs* 23 (2), March/April: 127-141.
- Rosenthal, Meredith B., et al. (Richard G. Frank, Zhonghe Li, and Arnold M. Epstein). 2005. "Early Experience with Pay for Performance. From Concept to Practice." *Journal of the American Medical Association* 294: 1788-1793.
- Rothstein, Richard. 2000. "Making a Case Against Performance Pay." *The New York Times*, April 26.
- Santora, Marc. 2005. "Cardiologists Say Rankings Sway Choices on Surgery." *The New York Times*, January 11.
- Schick, Allen. 2001. "Getting Performance Measures to Measure Up." In D.W. Forsythe, ed. *Quicker; Better; Cheaper: Managing Performance in American Government*. Albany, NY: Rockefeller Institute Press.
- Schiff, Michael. 1966. "Accounting Tactics and the Theory of the Firm." *Journal of Accounting Research* 4 (1), Spring: 62-67.
- Seidman, David, and Michael Couzens. 1974. "Getting the Crime Rate Down: Political Pressure and Crime Reporting." *Law & Society Review* 8 (3), Spring: 457-494.
- Simon, Herbert A. 1978. "Rational Decision-Making in Business Organizations." Nobel Memorial Lecture, December 8.
http://nobelprize.org/nobel_prizes/economics/laureates/1978/simon-lecture.pdf
- Skolnick, Jerome H. 1966. *Justice without Trial: Law Enforcement in Democratic Society*. New York: Wiley.
- Smith, Peter. 1990. "The Use of Performance Indicators in the Public Sector." *Journal of the Royal Statistical Society, Series A* 153: 53-72.
- Smith, Peter. 1993. "Outcome-related Performance Indicators and Organizational Control in the Public Sector." *British Journal of Management* 4 (3), September: 135-151.

- Smith, Peter. 1995. "On the Unintended Consequences of Publishing Performance Data in the Public Sector." *International Journal of Public Administration* 18 (2 & 3): 277 – 310.
- Stake, Robert E. 1971. "Testing Hazards in Performance Contracting." *Phi Delta Kappan* 52 (10), June: 583 – 588.
- Stecher, Brian, and Sheila Nataraj Kirby, eds. 2004. *Organizational Improvement and Accountability. Lessons for Education from Other Sectors*. RAND
http://www.rand.org/pubs/monographs/2004/RAND_MG136.pdf
- Steinbrook, Robert. 2006. "Public Report Cards - Cardiac Surgery and Beyond." *The New England Journal of Medicine*. 355 (18), November 2: 1847-1849.
- Thurlow, Martha. 2007. Personal correspondence, November 28ff, from Martha Thurlow, Director, National Center on Educational Outcomes.
- Timmins, Nicholas. 2005. "Blair Bemused Over GP Waiting Times." *The Financial Times*, April 30.
- Topol, Eric J., and Robert M. Califf. 1994. "Scorecard Cardiovascular Medicine: Its Impact and Future Directions." *Annals of Internal Medicine* 120: 65-70.
- Tu, Jack V., Kathy Sykora, and C. David Naylor. 1997. "Assessing the Outcomes of Coronary Artery Bypass Graft Surgery: How Many Risk Factors Are Enough?" *Journal of the American College of Cardiology* 30 (5), November 1: 1317-1323.
- Twigg, R. 1972. "Downgrading of Crimes Verified in Baltimore." *Justice Magazine* 5/6, June/July: 1, 15 – 18.
- UFT (United Federation of Teachers). 2007. *Report of the UFT Task Force on High Stakes Testing*. April 20. <http://www.uft.org/news/issues/reports/taskforce/>
- Uhlig, Mark A. 1987. "Transit Police Remove Officer For Quota Plan." *The New York Times*, December 21.
- U. S. General Accounting Office. 1994. *Health Care Reform. "Report Cards" Are Useful but Significant Issues Need to be Addressed*. September. GAO/HEHS 94-219. Washington, D.C.: General Accounting Office.
- U. S. General Accounting Office. 2002. *Workforce Investment Act: Improvements Needed in Performance Measures to Provide a More Accurate Picture of WIA's Effectiveness*. February. GAO-02-275. Washington, D.C.: General Accounting Office.
- Viadero, Debra. 2008. "Working Conditions Trump Pay." *Quality Counts, Education Week* 27 (18), January 10: 32-35.

- Vonnegut, Mark. 2007. "Is Quality Improvement Improving Quality? A View from the Doctor's Office." *The New England Journal of Medicine* 357 (26), December 27: 2652-2653.
- West, Martin. 2007. "Testing, Learning, and Teaching: The Effects of Test-based Accountability on Student Achievement and Instructional Time in Core Academic Subjects." In Chester E. Finn, Jr., and Diane Ravitch, eds. *Beyond the Basics: Achieving a Liberal Education for All Children*. Thomas B. Fordham Institute, pp. 45-62.
- Wilson, James Q. 1989. *Bureaucracy. What Government Agencies Do and Why They Do It*. New York: Basic Books.
- Wiseman, Michael. 2007. "Performing for Prizes: The High Performance Bonus as an Instrument for Improving Management of American Social Assistance." Paper presented at the 9th Public Management Research Conference, Public Management Research Association, Tucson, Arizona, 25-27 October 2007; October 22.

Acknowledgements

I am heavily indebted to Daniel Koretz, who has been concerned for many years with how “high stakes” can render test results unrepresentative of the achievement they purport to measure, and who noticed long ago that similar problems arose in other fields. Discussions with Professor Koretz, as I embarked on this project, were invaluable. I am also indebted to Professor Koretz for sharing his file of newspaper clippings on this topic and for inviting me to attend a seminar he organized, the “Eric M. Mindich Conference on Experimental Social Science: Biases from Behavioral Responses to Measurement: Perspectives from Theoretical Economics, Health Care, Education, and Social Services,” in Cambridge, Massachusetts, May 4, 2007. Several participants in that seminar, particularly George Baker of the Harvard Business School, Carolyn Heinrich of the LaFollette School of Public Affairs at the University of Wisconsin, and Meredith Rosenthal of the Harvard School of Public Health were generous in introducing me to the literatures in their respective fields, answering my follow-up questions, and referring me to other experts. Much of this paper results from following sources initially identified by these experts.

Access to literature from many academic and policy fields, within and outside education, was enhanced with extraordinary help of Janet Pierce and her fellow-librarians at the Gottesman Libraries of Teachers College, Columbia University.

Others have previously surveyed this field. Stecher and Kirby (2004), like the present effort, did so to gain insights relating to public education. But their survey has attracted insufficient attention in discussions of education accountability, so another effort is called for. Haney and Raczek (1994), in a paper for the U.S. Office of Technology Assessment, warned of problems similar to those analyzed here that would arise if quantitative accountability systems were developed for education. Two surveys, Kelman and Friedman (2007), and Adams and

Heywood (2007) were published or became available to me while I was researching this paper and summarized some of the same issues in a fashion which this paper, in many respects, duplicates. Susan Moore Johnson reminded me about debates in the early 1980s about whether teachers' intrinsic motivation might be undermined by an extrinsic reward-for-performance system.

A forthcoming Columbia University Ph.D. dissertation in sociology, contrasting “risk adjustment” in medical accountability systems with the absence of such adjustment in school accountability, should make an important contribution (Booher-Jennings, forthcoming).

This paper cites studies from the business, management, health, and human capital literatures, as well as previous surveys of those literatures, in particular Baker (1992), Holmstrom and Milgrom (1991), Mullen (1985), and Blalock and Barnow (2001). I am hopeful, however, that this paper organizes the evidence in a way that may be uniquely useful to education policy makers grappling with problems of performance incentives in education.

This paper has benefited from criticisms and suggestions of readers of a preliminary draft. I am solely responsible for remaining errors and misinterpretations, including those that result from my failure to follow these readers' advice. For very helpful suggestions, I am grateful to Marcia Angell, Julie Berry Cullen, Carolyn Heinrich, Jeffrey Henig, Rebecca Jacobsen, Trent Kaufman, Ellen Condliffe Lagemann, Lawrence Mishel, Howard Nelson, Bella Rosenberg, Joydeep Roy, Brian Stecher, and Tamara Wilder.

ENDNOTES (Citations to Bibliography)

-
- ¹ Simon, 1978, p. 352, 366.
 - ² Campbell, 1979, p. 85. See also citations in Darley, 1991; Baker, 2007; Koretz, 2007.
 - ³ West, 2007, p. 57.
 - ⁴ McMurrer, 2007; Jacob, 2005, p. 787; Dorph et al. 2007.
 - ⁵ NAEP, 1970, p. 81.
 - ⁶ Finn and Ravitch, 2007, p. 6
 - ⁷ Ridley and Simon, 1938, 1943, p.7.
 - ⁸ Ridley and Simon, 1938, 1943, p. 8.
 - ⁹ Ridley and Simon, 1938, 1943, p. vii.
 - ¹⁰ Ridley and Simon, 1938, 1943, p. 2.
 - ¹¹ Ridley and Simon, 1938, 1943, p. 47-48.
 - ¹² Ridley and Simon, 1938, 1943, p. 26.
 - ¹³ Ridley and Simon, 1938, 1943, p. 43.
 - ¹⁴ Ridley and Simon, 1938, 1943, p. 45.
 - ¹⁵ Ridley and Simon, 1938, 1943, p. 2, 16.
 - ¹⁶ Merton, 1957 (1968 Enlarged Edition), p. 253.
 - ¹⁷ Wilson, 1989, p. 117.
 - ¹⁸ U. S. General Accounting Office, 1994, p. 55.
 - ¹⁹ Smith, 1995, p. 284.
 - ²⁰ Holmstrom and Milgrom, 1991, p. 25-26.
 - ²¹ Nove, 1964, p. 294.
 - ²² Nove, 1964, p. 289.
 - ²³ Mullen, 1985, p. 165.
 - ²⁴ Mullen, 1985, p. 165.
 - ²⁵ Nove, 1958. p. 5.
 - ²⁶ Mullen, 1985, p. 165.
 - ²⁷ Mullen, 1985, p. 165.
 - ²⁸ Mullen, 1985, p. 166.
 - ²⁹ Alec Nove, cited in Mullen, 1985, p. 165.
 - ³⁰ Nove, 1958, p. 4; Litwack, 1993, p. 272-3.
 - ³¹ Blau, 1955, p. 38-42.
 - ³² Blau, 1955, p. 45-46.
 - ³³ Blau, 1955, p. 45-46.
 - ³⁴ Green, Passman, and Wintfeld, 1991, p. 1853.
 - ³⁵ Mullen, Rosenthal, and Frank. 2007, Slide 14.
 - ³⁶ Rosenthal, et al., 2004, p. 139.
 - ³⁷ Casalino, et al., 2007, p. 4.
 - ³⁸ Lu, 2007.
 - ³⁹ Smith, 1993, p.146-147.
 - ⁴⁰ Goddard, Mannion, and Smith. 2000, p. 141-142, 149.
 - ⁴¹ Smith, 1995, p. 291.
 - ⁴² Smith, 1993, p. 141-142.
 - ⁴³ Smith, 1993, p. 143.
 - ⁴⁴ Barnow and Smith. 2004, p. 249.
 - ⁴⁵ Barnow and Smith. 2004, p. 258-259.
 - ⁴⁶ Courty, Heinrich, and Marschke. 2005, p. 338.
 - ⁴⁷ Stecher and Kirby, 2004, p. 54, citing Courty and Marschke, 1997, p. 384.
 - ⁴⁸ Cited in Blalock and Barnow 2001, p. 505.
 - ⁴⁹ Heckman, Heinrich, and Smith. 2002, p. 808.
 - ⁵⁰ Courty, Heinrich and Marschke, 2005, p. 331, 341-342.
 - ⁵¹ Uhlig, 1987.
 - ⁵² Deming, 1986, p. 104.

-
- ⁵³ Moore, 2007. Jackman, 2004.
- ⁵⁴ Skolnick, 1966, p. 164.
- ⁵⁵ Murray, 2005.
- ⁵⁶ Johnson, Reiley, and Munoz. 2006.
- ⁵⁷ Campbell, 1979, p. 86.
- ⁵⁸ Booher-Jennings, 2006; Neal and Schanzenbach. 2007; Ballou and Springer 2007; Hamilton et al. 2007; UFT 2007.
- ⁵⁹ Rosenthal, et al. 2005, p. 1788-1789, 1793.
- ⁶⁰ Mullen, Rosenthal, and Frank. 2007, Slide 15.
- ⁶¹ Rosenthal, Fernandopulle, Song, and Landon, 2004.
- ⁶² Rosenthal, et al., 2004, p. 138.
- ⁶³ Bevan and Hood, 2006, p. 521.
- ⁶⁴ Rosenthal, et al., 2005.
- ⁶⁵ Smith, 1993, p. 149.
- ⁶⁶ Kelman and Friedman 2007, p. 19.
- ⁶⁷ Bird, et al. 2005, p. 20.
- ⁶⁸ Bevan and Hood, 2006, p. 531.
- ⁶⁹ Goddard, Mannion, and Smith, 2000, p. 149.
- ⁷⁰ Bevan and Hood, 2006, 523.
- ⁷¹ Timmins, 2005.
- ⁷² Bevan and Hood, 2006, p. 530-531.
- ⁷³ Bird, et al., 2005, p. 9, 20.
- ⁷⁴ Heinrich and Marschke, 2007, p. 26.
- ⁷⁵ Seidman and Couzens, 1974.
- ⁷⁶ Seidman and Couzens, 1974, p. 462.
- ⁷⁷ Seidman and Couzens, 1974.
- ⁷⁸ Morrissey, 1972; Twigg, 1972.
- ⁷⁹ Campbell, 1979, p. 85 (citations in text omitted).
- ⁸⁰ Butterfield, 1998.
- ⁸¹ Ridley and Simon, 1938, 1943, p. 3.
- ⁸² Ridley and Simon, 1938, 1943, p. 10.
- ⁸³ Ridley and Simon, 1938, 1943, p. 17.
- ⁸⁴ Ridley and Simon, 1938, 1943, p. 28.
- ⁸⁵ Stecher and Kirby, 2004, p. 104.
- ⁸⁶ Iezzoni, 1994, p. 4.
- ⁸⁷ Berwick and Wald, 1990.
- ⁸⁸ Berwick and Wald, 1990, p. 249; Iezzoni, 1991.
- ⁸⁹ Casalino, et al., 2007, p. 495.
- ⁹⁰ Green, Passman, and Wintfeld, 1991, p. 1852.
- ⁹¹ Krakauer et al., 1992, p. 330.
- ⁹² McKee, 1996, p. 430.
- ⁹³ Iezzoni et al., 1995.
- ⁹⁴ Tu, Sykora, and Naylor. 1997.
- ⁹⁵ Schick, 2001, p. 41.
- ⁹⁶ Associated Press, 1993.
- ⁹⁷ Kassirer, 1994.
- ⁹⁸ U. S. General Accounting Office, 1994, pp 40-41.
- ⁹⁹ U. S. General Accounting Office, 1994, p. 18.
- ¹⁰⁰ McKee and Hunter, 1994, p. 109.
- ¹⁰¹ McKee and Hunter, 1994, p. 110.
- ¹⁰² U. S. General Accounting Office, 1994, pp 5-6.
- ¹⁰³ U. S. General Accounting Office, 1994, p. 26.
- ¹⁰⁴ U. S. General Accounting Office, 1994, p 42.
- ¹⁰⁵ Roper et al., 1988, p. 1199.
- ¹⁰⁶ Epstein, 1995.

-
- ¹⁰⁷ Topol and Califf, 1994.
- ¹⁰⁸ Kassirer, 1994.
- ¹⁰⁹ Harris, Gardiner. 2007.
- ¹¹⁰ Steinbrook, 2006.
- ¹¹¹ Nakashima, 2007.
- ¹¹² Heinrich and Choi, forthcoming, Draft p. 17 and Appendix 1.
- ¹¹³ Heinrich and Marschke, 2007, pp. 21-23, 31.
- ¹¹⁴ Heinrich and Choi, forthcoming, Draft p. 29.
- ¹¹⁵ Wiseman, 2007; Horn, 2005.
- ¹¹⁶ Courty, Heinrich, and Marschke, 2005, p. 336-337; Heinrich 2004.
- ¹¹⁷ U. S. General Accounting Office, 2002, p. 9, 14.
- ¹¹⁸ Courty, Heinrich, and Marschke, 2005, p. 340, 342.
- ¹¹⁹ U. S. General Accounting Office, 2002, p. 28.
- ¹²⁰ Carnoy et al., 2005.
- ¹²¹ Jacob, 2005, p. 789-790.
- ¹²² Santora, 2005; Casalino, et al., 2007, p. 496.
- ¹²³ Santora, 2005.
- ¹²⁴ Bevan and Christopher Hood, 2006, p. 532.
- ¹²⁵ Altman, 1990.
- ¹²⁶ Green and Wintfeld. 1995.
- ¹²⁷ Altman, 1990.
- ¹²⁸ Casalino, et al., p. 496-497.
- ¹²⁹ Dranove, et al., 2003, p. 555-556, 577.
- ¹³⁰ Dranove, et al., 2003, p. 583-585.
- ¹³¹ Kerr, 1975, p. 773.
- ¹³² Courty, Heinrich, and Marschke, p. 338.
- ¹³³ Courty, Heinrich, and Marschke, p. 328.
- ¹³⁴ Dranove, et al., 2003, p. 581.
- ¹³⁵ Anderson, et al., 1991, p. 37, 39.
- ¹³⁶ Anderson, et al., 1991, p. 33.
- ¹³⁷ Anderson, et al., 1991, p. 45.
- ¹³⁸ Heckman, Heinrich, and Smith, 2002, p. 799.
- ¹³⁹ Barnow and Smith, 2004, p. 273, 278.
- ¹⁴⁰ U. S. General Accounting Office. 2002, p. 14-15.
- ¹⁴¹ Heinrich and Marschke, 2007, p. 15.
- ¹⁴² U. S. General Accounting Office, 2002, p. 16.
- ¹⁴³ Carnoy et al. 2005, pp. 29-33, 48-65.
- ¹⁴⁴ Kane and Staiger, 2002. (The version circulated in the summer of 2001 was an earlier draft of this published paper.)
- ¹⁴⁵ Kane and Staiger, 2002; Stake, 1971.
- ¹⁴⁶ Goldstein and Spiegelhalter, 1996, p. 405
- ¹⁴⁷ Campbell and Ross, 1968.
- ¹⁴⁸ Epstein, 1995.
- ¹⁴⁹ Kassirer, 1994.
- ¹⁵⁰ Green and Wintfeld.
- ¹⁵¹ Green and Wintfeld.
- ¹⁵² Bird, et al., 2005, pp. 14, 24.
- ¹⁵³ Farhi, 1996.
- ¹⁵⁴ McAllister, 1998.
- ¹⁵⁵ Hoyt, 2007.
- ¹⁵⁶ EPA. 1998.
- ¹⁵⁷ Kelman and Friedman, 2007, p. 10 and Figure 1.
- ¹⁵⁸ Ridgway, 1956, p. 241-242.
- ¹⁵⁹ Nove, 1958, p. 6.
- ¹⁶⁰ Blau, 1955 (rev 1963), p. 45.

-
- ¹⁶¹ Figlio and Getzler, 2002; Cullen and Reback, 2006; Jacob, 2005.
- ¹⁶² Figlio, 2005.
- ¹⁶³ Kantrowitz and Springen, 1997; Zlatos, 1994.
- ¹⁶⁴ Allington and McGill Franzen, 1992; Kantrowitz and Springen, 1997; Thurlow, 2007.
- ¹⁶⁵ Zlatos, 1994.
- ¹⁶⁶ Jaschik, 2007.
- ¹⁶⁷ Finder, 2007.
- ¹⁶⁸ Gormley and Weimer, 1999, pp. 148-149.
- ¹⁶⁹ Campbell, 1969, p. 415.
- ¹⁷⁰ Skolnick, 1966, p. 176, 181.
- ¹⁷¹ U. S. General Accounting Office, 1994, p. 38.
- ¹⁷² McKee and Hunter, 1994, p. 112; Smith, 1993, p. 148.
- ¹⁷³ Green and Wintfeld, 1995, Table 1.
- ¹⁷⁴ Green and Wintfeld, 1995; Epstein, 1995.
- ¹⁷⁵ Bird, et al., 2005, p. 7.
- ¹⁷⁶ Bevan and Hood, 2006, p. 529.
- ¹⁷⁷ Courty and Marschke, 2004, pp. 23-24, 28, 30, 33, 35, 49.
- ¹⁷⁸ Barnow and Smith, 2004, p. 269-270.
- ¹⁷⁹ U. S. General Accounting Office, 2002, p. 17.
- ¹⁸⁰ Barnow and Smith, 2004, pp 271-272, citing Courty and Marschke, 1997.
- ¹⁸¹ Gootman, 2007.
- ¹⁸² Bloomberg, 2007.
- ¹⁸³ Adams and Heywood, 2007, Tables 2 and 7.
- ¹⁸⁴ Roberts, 1989.
- ¹⁸⁵ e.g., Larkin, 2007, p. 9-10; Pfeffer, 1998, p. 115; Baker, Gibbons, and Murphy, 1994, p. 1125-1126.
- ¹⁸⁶ Patterson, 1992.
- ¹⁸⁷ Smith, 1990, p. 70.
- ¹⁸⁸ Ittner and Larcker, 2003, p. 89.
- ¹⁸⁹ Healy, 1985.; Jaworski and Young, 1992, p. 20.; Smith, 1990, p. 68; Schiff, 1966.
- ¹⁹⁰ Jaworski and Young, 1992, p. 20.
- ¹⁹¹ Myers, Myers, and Skinner, 2007.
- ¹⁹² Ridgway, 1956, p. 241, citing work by Chris Argyris.
- ¹⁹³ Jaworski and Young, 1992, p. 17.
- ¹⁹⁴ Oyer, 1998.
- ¹⁹⁵ Oyer, 1998, p. 156.
- ¹⁹⁶ Larkin, 2007, p. 3.
- ¹⁹⁷ Birnberg, Turopolec, and Young, 1983, p. 124.
- ¹⁹⁸ Adams and Heywood, 2007, p. 10.
- ¹⁹⁹ Ridgway, 1956, p. 247; Birnberg, Turopolec, and Young, 1983, p. 124.
- ²⁰⁰ Baker, Gibbons, and Murphy, 1994, p. 1125; Harvard Business School, 1984.
- ²⁰¹ Baker, 2002, p. 729.
- ²⁰² Holmstrom and Milgrom, 1991, p. 25.
- ²⁰³ Ittner and Larcker, 2003, p. 89.
- ²⁰⁴ Ittner, Larcker, and Meyer, 1997, p. 23-24.
- ²⁰⁵ Kaplan and Atkinson, 1998, p. 692-693.
- ²⁰⁶ Rothstein, 2000.
- ²⁰⁷ Bommer, et al., 1995.
- ²⁰⁸ Ittner, Larcker, and Meyer, 1997, p. 9.
- ²⁰⁹ Hamermesh and Biddle, 1994.
- ²¹⁰ Bommer, et al., 1995, p. 602.
- ²¹¹ Baker, 2002, p. 750.
- ²¹² Gibbons, 1998, p. 120.
- ²¹³ Pearce, Stevenson, and Perry, 1985, p. 263, 274; Baker, Gibbons and Murphy 1994, p. 1126.
- ²¹⁴ Rothstein, 2000.
- ²¹⁵ Gibbons and Murphy, 1990, p. 34-S; Baker, 2002.

-
- ²¹⁶ Pfeffer, 1998, p. 117.
- ²¹⁷ Gibbons and Murphy, 1990, p. 34-S.
- ²¹⁸ Holmstrom and Milgrom, 1991, p. 35.
- ²¹⁹ Deming, 1986, p. 76, 101-102. See also Pfeffer, 1998; and Deming Institute, 2007.
- ²²⁰ Kaplan and Atkinson, 1998, p. 368.
- ²²¹ Kaplan and Norton, 1996, p. 1-2.
- ²²² Schick, 2001, p. 50.
- ²²³ Kaplan and Norton, 1996, p. 220.
- ²²⁴ For a discussion, see Stecher and Kirby, 2004.
- ²²⁵ BNQP. 2007a.
- ²²⁶ BNQP. 2007b.
- ²²⁷ BNQP. 2007c.
- ²²⁸ Epperson. 2007.
- ²²⁹ Deci, 1971.
- ²³⁰ Deci, 1975, p. 219.
- ²³¹ Perry and Wise, 1990; Pfeffer, 1998, p. 116; Kreps, 1997, p. 360; Courty, Heinrich, Marschke, 2005, p. 323; Gibbons, 1998, p. 130.
- ²³² Perry and Porter, 1982, p. 94; Pearce, Stevenson, Perry. 1985, p. 262; Rainey, 1982, p. 288.
- ²³³ Crewson, 1997, p. 502-504.
- ²³⁴ Crewson, 1997, p. 504-505.
- ²³⁵ Cited in Perry and Porter, 1982, p. 90.
- ²³⁶ Rainey, 1982.
- ²³⁷ Crewson, 1997, p. 500.
- ²³⁸ Perry and Wise, 1990, p. 372; Kreps, 1997, p. 362-363.
- ²³⁹ Wilson, 1989, p. 60.
- ²⁴⁰ Goodlad 1984, p. 172.
- ²⁴¹ Vonnegut, 2007.
- ²⁴² Casalino, et al. 2007, p. 497.
- ²⁴³ Kelman and Friedman, 2007.
- ²⁴⁴ Goddard, Mannion, and Smith, 2000, p. 141; Bevan and Hood, 2006, p. 526-527.
- ²⁴⁵ Kelman and Friedman, 2007.
- ²⁴⁶ U. S. General Accounting Office, 1994, p. 56.
- ²⁴⁷ National Academies, 2008.
- ²⁴⁸ Heubert and Hauser, 1999.
- ²⁴⁹ Kerr, 1975.

Faculty and Research Affiliates

Matthew G. Springer

Director
National Center on Performance Incentives

Assistant Professor of Public Policy
and Education
Vanderbilt University's Peabody College

Dale Ballou

Associate Professor of Public Policy
and Education
Vanderbilt University's Peabody College

Leonard Bradley

Lecturer in Education
Vanderbilt University's Peabody College

Timothy C. Caboni

Associate Dean for Professional Education
and External Relations
Associate Professor of the Practice in
Public Policy and Higher Education
Vanderbilt University's Peabody College

Mark Ehlert

Research Assistant Professor
University of Missouri – Columbia

Bonnie Ghosh-Dastidar

Statistician
The RAND Corporation

Timothy J. Gronberg

Professor of Economics
Texas A&M University

James W. Guthrie

Senior Fellow
George W. Bush Institute

Professor
Southern Methodist University

Laura Hamilton

Senior Behavioral Scientist
RAND Corporation

Janet S. Hansen

Vice President and Director of
Education Studies
Committee for Economic Development

Chris Hulleman

Assistant Professor
James Madison University

Brian A. Jacob

Walter H. Annenberg Professor of
Education Policy
*Gerald R. Ford School of Public Policy
University of Michigan*

Dennis W. Jansen

Professor of Economics
Texas A&M University

Cory Koedel

Assistant Professor of Economics
University of Missouri-Columbia

Vi-Nhuan Le

Behavioral Scientist
RAND Corporation

Jessica L. Lewis

Research Associate
National Center on Performance Incentives

J.R. Lockwood

Senior Statistician
RAND Corporation

Daniel F. McCaffrey

Senior Statistician
PNC Chair in Policy Analysis
RAND Corporation

Patrick J. McEwan

Associate Professor of Economics
Whitehead Associate Professor
of Critical Thought
Wellesley College

Shawn Ni

Professor of Economics and Adjunct
Professor of Statistics
University of Missouri-Columbia

Michael J. Podgursky

Professor of Economics
University of Missouri-Columbia

Brian M. Stecher

Senior Social Scientist
RAND Corporation

Lori L. Taylor

Associate Professor
Texas A&M University

NATIONAL CENTER ON
Performance Incentives

**EXAMINING PERFORMANCE INCENTIVES
IN EDUCATION**

National Center on Performance Incentives
Vanderbilt University Peabody College

Peabody #43
230 Appleton Place
Nashville, TN 37203

(615) 322-5538
www.performanceincentives.org

