Northwestern | INSTITUTE FOR POLICY RESEARCH

RESEARCH EXCELLENCE • POLICY IMPACT

IPR Working Paper Series

WP-25-1

Why Strong Rewards in Government and Nonprofit Programs Don't Work: A Social Heisenberg Principle

Working Paper 1 in the Series: The Perils of Pay for Performance in Public Service Industries

Burton Weisbrod
Northwestern University and IPR

Version: January 10, 2025

Abstract

This paper explores the ways that many public programs incorporate well-intentioned funding and revenue incentives designed to improve service delivery, which unfortunately end up being undermined by “gaming” on the part of recipients. This result is sometimes illegal and sometimes not, but it is always economically inefficient and hence undesirable. Drawing on both historical and recent evidence, the paper examines how health care institutions, schools, higher education institutions, and a broad range of other public service industries derive financial incentives from performance outcomes regarding quantity or quality of service provided. This concept also applies to many government reimbursement programs for both nonprofit and profit-making service providers. Besides providing case studies illustrating distorted and undesirable outcomes, it also explains how and why those results occur, drawing on a “Social Heisenberg Principle” that relates similarities between the concept in theoretical physics and the concept that program performance measurement can and does change program performance outcomes. It ties this effect to reliance on “strong” financial incentives applied to simplistic performance measures that themselves distort outcomes in unexpected ways. It then discusses what should be done (or not done) to avoid the undesired outcomes.

About the Series: This working paper is the first in a series on "The Perils of Pay for Performance" for public service industries. The series highlights an important issue today, which is how for-profit firms, nonprofit organizations, and governmental agencies can coexist in many parts of a modern economy, with each playing a role supporting "better" performance. Other papers in the series delve deeper into those issues for specific industries including K–12 schools, health care, and higher education.

Acknowledgements. This research and case study evaluation was assisted by many people over the course of a decade of work. I am particularly grateful for important editorial assistance and discussion refinement provided by Northwestern University's Dr. Evelyn Asch, now retired, and by Dr. Rachel Golden, a free-lance copy editor. I also thank the many students at Northwestern University who served as research assistants: Chung-Hyun Kim (courts), William Russell (education), Grant Johnson (K–12 education), Layne Kirshon (museums, libraries, policing, higher education), Travis Howe (charities), Carol Shou, Candice Choi, and Nina Munoz (health care), Camille Liu (museums and K–12 education), Grace Dennis (performance measurement), plus Andrew Ruth, Tyler Goff, Tory Do, Rui Du, Yushi Luo, and Robert Eli Winter (higher education). All conclusions and opinions are exclusively those of the author.

INTRODUCTION

"Pay for Performance" is a term widely used in business for employee pay that is tied to achievement of performance goals. The term is also used in the health care industry for hospital reimbursement schemes that are tied to patient care and resource use outcomes. But the concept can also apply for health care institutions, schools, higher education institutions, and a broad range of other public service industries that all derive financial incentives from performance outcomes regarding quantity or quality of service provided. Indeed, this same concept applies to many government reimbursement programs for both nonprofit and profit-making service providers.

This paper explores the ways that many public programs and policies incorporate well-intentioned funding and revenue incentives designed to improve service delivery, which unfortunately end up being undermined by "gaming" on the part of some recipients. It discusses how and why pay-for-performance (P4P) in the public policy realm sometimes works well, but sometimes fails badly, and the conditions that lead in each direction are predictable. Moreover, the unwanted effects can be avoided or at least reduced, but there will be costs. The challenges of balancing the favorable effects of offering "strong" – that is, "large" -- rewards is that they act as incentives to increase "employee" (service provider) income even when the "employer" (society) is not made better off, while avoiding the unfavorable effects (costs) of deceptive actions that hide bad results, are examined in this paper and in subsequent papers in this series – covering industries such as schools, higher education, hospitals, prisons, museums, courts, and more. Some date back to the writing of the U.S. Constitution in 1789.

The degree of success or failure in linking the *pay* that anyone receives, for better or worse *performance,* hinges on two dimensions of measurements: one is their *accuracy*; measures that avoid both exaggerations and understatements of true performance, which requires that *all* forms of performance be calibrated, not just those that are easy to observe, such as the death rate at a hospital, the number at a secondary school, the average student score on a state's standardized mathematics test for a specific school level, or, at a museum; and (2*) valuations* of each form of measurement, to permit them to be combined into a *total* contribution to the employer, thereby *aligning* each worker's compensation with his or her performance productivity. That process would establish incentives for each worker to be efficient.

But those are often not small tasks. And unless these conditions *all* hold, and simultaneously, P4P will fail, often producing the opposite of its intended goals. Hence, the "perils."

Government and private nonprofit organizations can face a multiplicity of goals, compared with private firms that can be expected to pursue maximum profits for their shareholders. This can create obstacles to unbiased measurement of goal achievement that are especially troubling for government and private nonprofit organizations. To see why, I will explore a wide range of industries and jobs, showing the diversity of obstacles to measuring, valuing, and rewarding their "performance". An important lesson emerges: any one approach, such as always striving for *stronger* performance-based

rewards, or always relying on a particular organization form – a private firm, government authority, or private nonprofit – is not desirable, it is *counter*-productive.[1]

Imagine an activity in which the problems of measuring, valuing, and rewarding true performance, and doing it accurately, without systematic bias, are so e that employers have virtually abandoned efforts to link compensation to *any set* of performance measures.

Implausible? No.

For instance, we can look to examples of intentional and persistent use of extremely *weak* incentives. One is federal judges. Encompassing District Courts, Circuit Courts, and the U.S. Supreme Court, these judges are not rewarded for performing "better" as they hold lifetime appointments subject to termination only by formal legislative impeachment and conviction, or the judge's resignation or death their annual salaries are rigidly fixed -- by Congress – and while those salaries are changed occasionally, they are equal for all judges at each court level, independently of judge's performance, productivity, experience, duration in office, or other judges' appraisals. And it has remained this way since the founding of the court system in 1789. Why?

To this day there has been essentially no attention to the Pay-for-Performance (P4P) process for federal judges – not to how to *measure* each form of performance better, not to how to *value* the measures, and not to ways to *link* the measures to a judge's *compensation*.

The weak rewards seem extremely inefficient, but to reach a different conclusion; there is indeed an efficiency-based case for using even extremely weak rewards under particular conditions, and those conditions do, in fact, exist in many parts of a modern economy, including not only federal courts, but also schools, hospitals, prisons, charities, museums, and more.

Regardless of the industry or activity, the root issues are the same: how "serious" are the unintended effects of developing, instituting, and sustaining a P4P system that meets three conditions: it *does reward* forms of performance that are easy to measure and value, it does *not* reward forms of performance that are costly to measure and value, and it avoids encouraging gaming the compensation system by rewarding what *appears* to be "better" performance, but actually is not.

There is a lesson: It is neither efficient nor equitable to do the "best we can," rewarding what can be easily and accurately measured and valued, but disregarding other effects. The resulting performance measures and rewards are systematically *mis*measured and biased -- either over- or under-rewarding true performance. These *mis*measurements are predictable, and their rewards are inefficient – either excessive or inadequate. the forms of performance that are both easy to achieve and more highly rewarded will be concentrated upon by providers, while the less well-rewarded elements of performance receive scant attention, if any. Here are a few case examples:

▪ ***Hospitals.*** Quality of hospital care is clearly relevant to how performance should be measured and compensated. But *quality* and *quantity* of hospital care are not well defined and valued, and so they *can* be gamed through manipulations of performance measures if that would generate greater rewards for "better" performance by specific stakeholders. There is no way to entirely evade these issues, but

[1] For an insightful examination of pay-for-performance mechanisms and challenges, see Kreps, *Motivation Toolkit*.

recognizing the differential effects of relying on strong and weak rewards under identifiable conditions is a valuable launching point that addresses the incentives for gaming. Here are two examples.

▪ ***Schools.*** It is clear why teachers devote more time and effort to material they believe is more likely to be covered on standardized tests and, because higher student test scores are likely to yield different rewards to the students, teachers, and administrators involved in the education process, self-interest forecasts that inflation of exam grades and course grades will incentivize teachers and administrators to concentrate on subject matter that is likely to bring more acclaim, promotions, and other rewards to schools, students, principals, teachers, and administrators. Evidence of gaming by school administrators, teachers, and students in a number of large American cities will be examined later, highlighting the variety of gaming opportunities in legal, illegal, and borderline forms.

▪ ***Police Departments.*** Why expect police to devote time and effort to reducing minor law violations such as jaywalking or used-bicycle thefts, which the FBI has judged to be not serious enough to warrant reporting in its monthly crime statistics, rather than to concentrate on reducing the numbers of more serious, "Part I," offenses, which include homicides and new car thefts, and are used to measure and reward police department *performance* and its changes over time. Insofar as police P4P systems reward *reductions in the more serious crime rates*, and larger reductions are rewarded more, police have stronger incentives to under-report less-serious violations that would have little or no effect on performance-based compensation anyway, and concentrate on the more highly rewarded accomplishments. Police thus had the incentive, for example, to understate the number of their car "stop and frisk" actions that involved racial and other minorities. The goal was to intentionally report to supervisors' fewer stops of racial minorities than were actually made; by *understating* the true number of police arrests of "suspicious" drivers who were, police were understating their true reliance on minority profiling to provide false evidence of reduced police profiling and improved performance.

But in at least one large U.S city the police patrols were simply manipulating the performance reports, by adopting a deliberately falsified *measurement* methodology. They cheated in their reporting.

Evidence of numerous forms of gaming throughout the economy lie ahead, in a diverse set of industries.

Performance Measurement and Mismeasurement: Keys to Efficiency and Lessons from the Provena Covenant Medical Center Case

The central issue permeating P4P systems is measurements – how to make them comprehensive, not partial, and unbiased, not consistently under- or over-counted or valued. But how to deal with those obstacles? How should the "values" of better performance in any industry be quantified and valued when the performance measures are not determined simply by competitive market prices, but are sold at a number of government-subsidized prices, or are given to the poor at a price of zero, are offered at varied prices to consumers at peak and off-peak hours and days, or are financed by voluntary donations of time and commodities? With a multiplicity of organization goals, output prices, performance measures and rewards, strong P4P incentives are particularly problematic.

The answers to at least some of these measurement questions turned out to be critical in the March 2012 case of the Provena Covenant Medical Center, an Illinois nonprofit hospital that had tax-exempt status under federal and some state tax laws, including real-estate taxes. But then the automatic state and local tax exemptions ended abruptly, as state law was reinterpreted to require that

to retain its property-tax exemption, a nonprofit hospital (only nonprofits were singled-out for attention) had to give evidence that it was providing "charity care" that exceeded in its "cost" to the hospital the dollar value of the property tax the state was foregoing. Only then would the property tax bill be waived, for that year. Only if the "dollar amount" of a nonprofit hospital's performance in providing c*harity care* for the poor and uninsured *was* reported to the state's Department of Finance to have exceeded the hospital's property tax bill would the tax bill be waived.

But how was the "amount" of charity care a hospital actually provided supposed to be measured and valued? How and by whom was the measurement process to be monitored, to assure accuracy and honesty, rather than some form of statistical gaming?

The state did have some answers, but they left substantial latitude for a nonprofit hospital to game the state's rules. Millions of dollars of taxes could be at stake for a hospital even in a single year. And since each nonprofit hospital reported for itself, it exercised considerable discretion over its choices of accounting techniques for calculating its charity care and whether it reached the property tax exemption threshold.

Provena Medical Center had reported spending $831,724 on charity care for the first year in contention, 2002 -- less than the nearly $1.1 million of property tax the state would lose if it granted the property tax exemption. But the State did not lose the case, and the Illinois Supreme Court upheld the state's position.

Beware of Self-Reported Performance Data

The State government's reporting process required each nonprofit hospital to disclose the dollar amount of *charity care* it provided, a process that was both common and dangerous. It was an invitation to over-report its charitable activities, particularly if the hospital's normal reporting procedure brought it "close" to the tax-exemption lower boundary. Self-reporting gave a hospital discretion in reporting its own charity-care and thereby reduce its property tax liability. The door to gaming the P4P incentive system opened wider, despite the dampening effects of government audits.

Self-reporting was just the beginning of a gaming process involving many stakeholders, including governments, taxpayers, consumers, and not only hospitals but many other industries, particularly with sizable nonprofit components. Each participant in the tax-collection process had a stake, although the basic challenges were typically how to retain as much of their *revenue* as possible, and how to shift as much of their *costs* as possible to other stakeholders. In the case of charity-care services, as seen by the Illinois Department of Finance, the issue was how to incentivize nonprofit hospitals to care "sufficiently" for the poor or lose their entire property-tax exemption.

The court ruling ended a nonprofit's property tax exemption unless the hospital provided charity care that was a net money-saver for the state, costing it less in foregone property tax revenue than it would incur to pay for the charity care directly.

Quid pro Quo Financing Could be Used for Other Public Services

From any state's long-term finance perspective, more significant than the struggle over nonprofit hospitals' exemptions from property taxation was and is the potential for states to expand *quid pro quo* relationships well beyond nonprofit hospitals and charity care. A government agency can barter with any service provider, not only hospitals, offering tax reductions in return for increased provision of specific services or at lower prices to particular consumers. This hospital case was just one version of the

use of tax policy, for financing one industry, hospitals, one state, Illinois, and one form of output performance, care for the indigent. But there are countless other opportunities to expand such *quid pro quo* trades through which governments could exchange tax benefits for provision of public services.

Later I will highlight many opportunities for service suppliers in a wide range of industries, to *game* the rules linking performance to rewards. All it takes are manipulations of the *reported* charitable, public-good, activities to the value of achievements. The manipulations are the *Perils of P4P*.

But first a brief digression to explain my reference to a *Social Heisenberg Principle,* and its relevance to performance measurements, valuations, and incentives. There is, it turns out, a striking analogy between physicist Heisenberg's *Uncertainty Principle*, in the field of quantum mechanics, and my attention to the ways in which, in the economic and other social science realms, the *measurements* of performance and the *rewards* for their advancement are linked. With multiple goals, some will conflict, distorting incentives.

Why" distorting'? Some goals are easy to measure, value, and so to reward. Rewards act as magnets attracting effort and increasing productivity. In the process, though, they crowd-out attention to other, less easily observed, social goals and the effects of inattention to them.

Other papers in this series show the scope of resulting conflicts for specific industries. But first, a brief history of science will help to connect the evolution of P4P in a modern economy to developments in the world of physics and specifically its attention to quantum mechanics and the behavior of minute atomic and sub-atomic particles. The two worlds of incentivizing greater productivity and explaining atomic and sub-atomic particles' responses to measurements are very different --and very alike.

Performance Measurement Meets a Social Heisenberg Principle

In 1927 a 26-year-old German theoretical physicist, Werner Heisenberg, presented a paper on what came to be known as *Heisenberg's Uncertainty Principle*. His research involved a contentious theoretic debate question in quantum mechanics: was it logically possible to measure two related goals simultaneously – specifically, the *location* and the *motion* of a *subatomic* particle such as an electron? No, he concluded – a view still prevailing nearly a century later.[2]

My target, by contrast, is *macro*scopic – are there limitations to using performance-based incentives in an industrialized economic system with multiple goals that conflict? Both the microscopic and macroscopic perspectives address a similar issue – are there unavoidable side-effects of pursuing multiple goals that conflict? Yes, there are.

Whether the Heisenberg Uncertainty Principle had an analogy in the wider social science world was beyond his scope. He explained that measuring one dimension of an atomic particle's behavior alters the measurements of others. Today, physicists distinguish Heisenberg's "uncertainty principle," which focuses on what can be measured, and an observer effect," which focuses on how mere observation can affect behavior of atomic or subatomic particles.[3] By extending an analogy of the latter concept to human behavior, one might say that if an anthropologist travels to a remote place to study

---

[2] American Physical Society, "Heisenberg's Uncertainty Principle."
[3] Salkind, *Encyclopedia of Research Design.*

how its people live, the very act of an outsider watching could change their actions, thereby inducing observer effects; measuring can alter the measurements.

These forces provide links to my focus in this series of papers on *The Perils of Pay-for-Performance.* The perils result largely from the use of strong, high-powered, rewards for performance that are *easy* to measure and value, even if that diverts incentives away from other forms of performance that are more *costly* to measure and so are essentially overlooked or cast aside, treated as worthless. Rewarding some forms of performance, but not others, or rewarding them differently, are recipes for distorting incentives.

In this first paper, I highlight the systematic, non-random, dangers resulting from using strong rewards for measured performance, particularly in government agencies and private nonprofits. But why only those organizations, and why not private, for-profit, firms? Because government and nonprofit organizations are more likely than private firms to produce outputs that are costly to measure fully, to value, and to reward. The result is that using strong rewards in those realms is especially problematic; some types of good performance are assessed, but not others, and some forms of poor performance are penalized, but not others. Measurements can be counter-productive, not merely perilous.

Today, the challenge to using stronger incentives based on measured performance, regardless of the industry, product, or service involved, remains tethered to measurements and their valuations, even when performance has multiple and conflicting dimensions. Efficient incentives call for linking rewards to *total* performance, which requires identifying, measuring, and valuing *all* forms of performance. While personnel typically have key knowledge of their own performance, it is often not in their self-interest to reveal it, especially if they would be penalized for having overstated their performance when it was costly for other stakeholders to detect.

Do Efficient Incentives Differ Among Ownership Forms? Choosing Among Forms

Inefficient incentives are especially costly to avoid when they result from information that one party to a transaction has but others do not – a common asymmetric-information situation.

Information asymmetries (inequalities) cannot be cut, though, let alone eliminated, simply by shifting production from one ownership form to another, such as by privatizing the U.S. Postal Service (USPS). True, private firms, government agencies, and nonprofit charities do have different advantages and disadvantages because of taxation and regulation. These encourage the view, though largely erroneous, that, for example, the alleged inefficiencies of the USPS can be corrected easily by converting it to another ownership form.

But not so fast. Each ownership form has strengths and weaknesses, which translate into advantages and disadvantages in the pursuit of particular goals and opportunities.

Private firms have the widest range of opportunities with the fewest legal restrictions. They are generally free to decide what to produce, how to produce them – that is, what combinations of labor and capital to use in production -- how to incentivize employees to be more efficient, how to establish values in uncompetitive markets, and how to distribute any profits.

Governments and nonprofits are subject to other constraints. Some are determined by their goals, which may differ from those of private, presumably profit-oriented, firms, and some from governmental regulations, greater access to volunteer labor and private monetary donations; they, too, may differ from private firms and some forms of government agencies and nonprofits. A non-distribution

constraint (NDC) also limits nonprofit organizations' opportunities and incentives to grow, to enter new product markets, exit some current markets, and raise capital by selling ownership shares, because some types of nonprofits are not legally permitted to have private owners, nor to distribute to their trustees, management, or employees what would otherwise be taxable profit.

Another feature of some outputs that affect consumers' preferences for dealing with one or another form of supplier organization, and hence affect sellers' choices of their own forms -- and they do have choices, with differing advantages and disadvantages – involve control of output "quality." In higher education, for example, class size is generally easy to observe, but whether it affects "learning," whether online courses offer lower quality, and how quality *should* be measured, are far from resolved. Students and parents often make such judgments, of course, but the fact remains that higher education has many attributes that can be easily gamed; for example, faculty can be encouraged or discouraged from inflating course grades, increasing graduation rates, and devising ways to improve a school's ranking in the *U.S. News & World Report* annual college evaluations. I turn to gaming on higher education in a separate paper, highlighting imaginative applications of rules and restrictions that generate better *measured* performance, even if often not achievement of hard-to- measure and value contributions to educational goals.

Rules, laws, and other restrictions or mandates on behavior of any form of organization are of little effect, though, without their enforcement. The NDC restricts nonprofit organizations, but not-for-profit firms, in important ways, yet enforcing those restrictions poses costly hurdles, even for the IRS, the federal tax collector. Congress provides the IRS with a budget earmarked for enforcing restrictions on tax-exempt nonprofits, and it regulates taxpayer incentives to influence accounting practices and personal income tax deductions for charitable donations under section 501(c)(3) of the Internal Revenue Code. The funding also supports detection of nonprofits' efforts to hide taxable income. The IRS, however, has no incentive to maximize its profits, nor even the Treasury Department's profits, from tax collections, for any IRS profits go to the Treasury Department, and are not available to the IRS or its employees. Industrialization and advances in science have ushered-in conditions in which new forms of products, services, and information inequalities between buyers and sellers were emerging. These inequalities undermined buyers' confidence in loosely regulated or unregulated private markets. The greater the inequalities of information holdings, the more potentially efficient is the shift of consumers who are aware of their information handicaps, away from private firms and to greater reliance on government or private nonprofits. Those suppliers are more restricted in their incentives to advance their self-interests by expanding their information superiorities over consumers. That was the path taken by the VW Automotive Group, and it worked, temporarily.

Information Asymmetries (Inequalities) and the Undermining of P4P
-- Lessons from Used-Car and Blood-Transfusion Markets

There are large segments of the economy where informational asymmetries are fundamental obstacles to using strong P4P rewards, and to relaxing regulatory restrictions on private firms. These parts of the economy, and their implications for public policy, are my primary focus.

A half century ago, in 1970, two seemingly unrelated research publications appeared. They were very different, and very alike. Both emphasized the inefficiencies resulting from information inequalities between sellers and buyers. One, however, was written by an American economist, George Akerlof, who later became a Nobel Laureate in economics; the other by a British social work researcher, Richard

Titmuss. Both publications showed the adverse effects of information inequalities between buyers and sellers, even in diverse countries and markets.

Akerlof highlighted the adverse effects of information inequalities in the American market for second-hand goods such as used cars, although he made it clear that information asymmetries handicapped many other markets.[4]

Titmuss directed attention to differences between markets for human blood transfusions in the U.K. and the U.S. He focused on the greater prevalence of hepatitis-B infections in the U.S., attributing the difference to greater information inequality between suppliers and demanders of blood used for transfusions in the U.S.; His theme was captured by the title of his book, *The Gift Relationship,* He saw the greater prevalence in the U.S. of tainted blood transfusions carrying the hepatitis-B virus as explainable by informational inequalities between blood donors and recipients; at the time of his research in the 1960s the key in both countries was the exchange of blood for money; when blood *donors* knew more than blood *recipients* about the risk of receiving tainted blood in a transfusion – especially in the U.S. -- and there was no test available at that time to close the information gap, so disease transmission would be more likely follow to follow, as he emphasized.[5]

Titmuss claimed that nearly all blood *donations* in the U.K. were actually unilateral *gifts*, provided through the British National Health System (NHS) and without payment to a blood donor. [6] In the U.S., by contrast, he saw the bulk of blood supplied as effectively *sales*, not gifts, donors being compensated in cash or other goods such as "free" future transfusions if needed. Apparently, under the NHS the supply of blood for transfusions was sufficient to meet the "need" without compensation, unlike in the U.S.

Titmuss' interpretation was that American blood donors were offering their blood largely to raise money to finance their drug addictions. They knowingly engaged in needle-sharing and other unsafe practices, he argued, and were selling their often-tainted blood to satisfy their addictions. Not so for British donors, who were portrayed as motivated simply to help their countrymen and women by donating blood at a price of zero; there was no financial incentive to *sell* one's blood when anyone in "need" knew it could be obtained *free* from the NHS, with little or no safety concern.

In the U.S., by contrast, a patient typically not only had to pay for a blood transfusion, but was less confident of its quality, which may have been compromised by the unsafe actions of drug-addict donors. It is also noteworthy that in the late 1960s, when Titmuss was finishing his book, *The Gift Relationship*, knowledge of how to test a prospective blood donor for hepatitis-B was quite limited; the time required for the blood-safety test exceeded the blood's shelf-life, making the test virtually useless.

When Akerlof highlighted the adverse effects of information asymmetries between sellers and buyers he turned not to a government agency, but to a private market – for used-cars. He reasoned that if a car owner considered selling, and if the owner knew that the car was in exceptionally good condition, he might well expect a disappointing price that did not reflect the car's outstanding condition

---

[4] Akerlof, "The Market for Lemons"
[5] Titmuss, *The Gift Relationship*
[6] Titmuss, *The Gift Relationship*

and value; buyers, however, at that time were typically unaware of its true value, so it was wise to keep the car off the market, rather than "underprice" it, selling it at a price around the *average* for cars of that make, model, mileage, vintage, and other observable characteristics. But neither could the owner easily convey evidence of its unobserved high quality through traditional metrics then available at low cost to buyers. So, the car's *truly excellent* quality would not be rewarded, as the high-quality end of the used-car market collapsed under the weight of the deficient quality-measurement technology of the 1960s.

Information differences among stakeholders are central to the industries and commodities on which I concentrate. They explain, for example, why hospitals, k-12 schools, higher education, policing, jails, museums, philanthropies, and federal courts have *small* presence in the private enterprise sector but large ones in government and nonprofit sectors.

Why? Because those industries impose challenging information requirements for measuring, valuing, and rewarding or penalizing performance. And strong, high-powered, incentives for better performance entice more-sophisticated gaming of rewards and penalties.

While strong rewards present incentives to game the reward system, *weak and low-powered,* rewards for measured performance have their own handicaps. They thwart incentives to be more productive, instead encouraging actions that often do not stimulate effort and productivity.

Both Akerlof and Titmuss had recognized that these information-based incentives were promoting inefficient gaming, in their independent analyses of behavior in the U.S. and British markets for used cars and human blood. In the analyses ahead I will show many other illustrations of intentional gaming practices in industries where incentives were unintentionally being distorted, often in complex ways such as inflating *students'* standardized test scores in k-12 *schools*, exaggerating *hospitals*' provision of "charity care," and encouraging *police* who operate patrol cars, to misreport the comparative frequencies of their stopping cars with Black drivers relative to White drivers who were allegedly driving "suspiciously." Police on routine patrols were aware of the pressure on them to avoid racial or ethnic "profiling," but were also aware of their informational advantages over their supervisors in knowing whether their reports were truthful and accurate, and when incentives encourage the converse. In industry after industry, I will return to these and other illustrations of gaming in forms that are responses to *unintended* incentives for better-informed stakeholders to reap substantial rewards based on *measured performance,* even when *true* performance is small or negative.

By contrast, when *weak* rewards prevail, they typically *under*-reward forms of performance that are costly for supervisors to measure and value, yet at the same time they avoid *over*-rewarding forms of performance that are easily measured and valued. So rewards that are only weakly linked to performance can be expected to dominate segments of an economy where gaming of performance-measures is costly to prevent but easy to exaggerate, thereby attracting stakeholders with private informational advantages.

Later in this paper, I will turn from a focus on the causes and consequences of adopting *strong* rewards for measured performance, to the opposite end of the incentive spectrum, the incentive effects of *weak rewards.* Why would any employer, public or private, choose to adopt a reward system that intentionally detaches an input provider's rewards from his or her productivity? And why would the detachment of rewards from measured performance be sustained?

Odd? indeed. Yet it characterizes the market for federal judges in the U.S. today. It is mysterious why any employer would adopt a P4P reward system that essentially proclaims that there should be no incentives, no differential rewards or penalties, for "better" or "worse" judicial performance. Once a federal judicial candidate has been nominated by the President of the U.S. and confirmed by a vote of the U.S. Senate, and officially sworn-in the judge's compensation is essentially fixed by Congress but for a possible impeachment by the House of Representatives and conviction by the Senate, or the judge's death, voluntary retirement, or a rare appointment to a higher-level court.

Moreover, federal judicial appointments are lifetime, subject to these limited exceptions. And all judges at a given court level – district, circuit, or Supreme Court -- are paid essentially equally, regardless of their duration of service, age, or other proxy measures of the "quality" of their judicial performance.

What is most notable, overall, is less the extremely *weak* reward structure of federal judicial compensation, than its longevity. While Congress has approved across-the-board salary increases for federal judges at a particular judicial level, the absence of connections between a particular judge's measured performance and compensation – that is, the weak rewards for performance --has been sustained since its adoption in the U.S. Constitution in 1789. The weak but durable reward structure for a federal judge's performance appears to be no accident or oversight; there are virtues as well as shortcomings to diverse incentives under identifiable circumstances; I will return to the nature of the circumstances that favor strong or weak rewards.

The link between measured performance and its compensation has withered through time, as new forms of gaming have evolved. While the lifetime judicial appointments to federal courts have continued, now in their 233rd year, they now amount to a considerably longer period of *expected* employment, because adult life expectancy and working-life expectancy have increased markedly. Between 1850 and 2020, for example, a federal court appointment became increasingly valuable as working-life expectancy increased dramatically; for example, a 50-year- male's (lie expectancy increased by 35 percent, from 21.6 years to 29.2; and for females the increase was more than 40 percent, from 23.5 years to 33.0 years.[7]

Even more striking is that apart from the monetary value of the increasing longevity and earnings in the context of a lifetime employment appointment, the remarkably weak P4P compensation structure for all federal Judges has not been changed in over 233 years, since adoption of the U.S. Constitution in 1789, although it was not until 1992 when the Constitutional amendment regarding judicial compensation was ratified by three-fourths of the States. The Constitution and its Bill of Rights – the first ten amendments to the Constitution --have been amended 27 times, as recently as 1972, but has not differentiated among judges at a particular judicial court level, unless he or she was either impeached and convicted, appointed to a higher court, died, or chose to retire – all rarities.

Have the centuries of weak, virtually zero, rewards for "better" federal judicial performance simply been a long series of mistakes? Were repeated errors not corrected even though dozens of other Constitutional amendments were being made? Unlikely, although a sitting federal judge with whom I have spoken had no doubt; he thought the lifetime judicial appointments and rigid compensation

---

[7] 2020 data from: NVSS, "Provisional Life Expectancy Estimates…", p.2; 1850 data from National Vital Statistics as cited in InfoPlease, 2012.

structures were *not* mistakes; rather, in his judgement the errors were the failures of *state* and *local* governments to emulate the federal judicial reward structure and its guarantees.

Today, federal judges' employment security and extremely weak incentives for better measured performance have been sustained, continuing to include *lifetime* job tenure, subject only to formal impeachment and conviction, resignation, or death. Compensation continues to pay all federal judges at a specific court level equally, regardless of their length of service, age, or measured "quality" of their judicial decisions. These ultra-weak incentives still apply to federal courts, and not to state or municipal courts -- nor to other industries or occupations. The extraordinary, multi-century, persistence of such weak rewards for federal judges implies that their continued application in this realm has been judged to be more efficient than the stronger rewards typical of the bulk of the economy where performance is more easily observed and more fully measured and valued. There are signs, though, of growing political resistance to retaining the *status quo* of judicial compensation that is unconnected from measures of a judge's performance.

Yet the fundamental challenges to P4P remain how to measure, value, and then reward performance while avoiding distortionary gaming that causes goals to conflict. Heisenberg had emphasized a century ago, in developing his Uncertainty Principle, the *impossibility* of *simultaneously* measuring both the speed and trajectory of a sub-atomic particle's motion. In my P4P measurement-system context, there are also two likely competing goals, that generate defective measurements as they are pursued: performance measures are often i*ncomplete*, omitting at least some components of performance *quality*, and they are often *biased*, either consistently under-estimating or over-estimating the true measurements or rewards.

The results are clear it is costly, if not unfeasible, to measure performance, to value the measures, and to reward and thereby incentivize the agents and decisions responsible for the biased reported productivity. To be efficient, performance measures must be fully and accurately gauged and rewarded -- formidable requirements. But unless all relevant effects, favorable and unfavorable, are rewarded and penalized, the P4P process will fail, as incentives misdirect resources.

Heisenberg had recognized a century ago, the problems associated with multiple goals and the interdependencies of their joint pursuit. In the world of quantum physics, he concluded that the pursuit of multiple goals, specifically to measure simultaneously the location of an atomic particle and its trajectory, was not merely difficult; it was impossible. The act of measuring location, he concluded, would alter the trajectory, which is now termed the "observer effect." The conflict of goals that became known as the *Heisenberg Uncertainty Principle* remains a guiding light in physics, linking, at the theoretic level, the *micro*scopic concepts of quantum physics and the *macro*scopic perspective in the world of economics.

The complex and often countervailing incentive effects of multiple goals will resurface when we examine the applications of P4P in settings such as Veterans Administration (VA) Hospitals, k-12 public schools, colleges, police, and jails that are owned by state Departments of Prisons but operate under contracts with private for-profit firms. Whatever the industry, though, all P4P reward systems are designed, or at least from a social efficiency perspective, should be designed, as instruments to enhance "productivity," to cut "costs," and to discourage gaming of incentive systems.

Every form of organization ownership faces advantages and disadvantages. There are subsidies and taxes, restrictions, prohibitions, and mandates, that impose both similar and differential constraints and that narrow some options and expand others. For example:

- P*rivate firms* are generally not limited in the products, services, and pricing policies they employ in pursuit of profit, subject to general restrictions on safety, anti-trust law violations, and other regulatory restrictions, but their profits, if not derived from their tax-exempt activities, are taxed as personal or corporate income.
- G*overnments* and private *nonprofits,* however, are generally *not* taxed on profits from provision of "collective" ("public") services for which they receive tax-exemptions and subsidies.

But these and other restrictions and opportunities bring ambiguities and encourage gaming.

Choosing Among Organization Forms: the Non-Distribution Constraint (NDC) and Other Differential Restrictions Among Government, Nonprofit, and For-Profit Producers

Many activities of *government* and private *nonprofit* organizations are constrained by tax laws, subsidies, and other legislation, but *for-profit firms* are more privileged. They are allowed to incentivize their employees and managers to be more productive in advancing the firm's goals, traditionally assumed to be profit-maximization. Government agencies and private nonprofit organizations are allowed, though, to pay performance-based bonuses to employees and managers, *if* the bonuses are *not* based on what would otherwise be treated as taxable "*profit."*

Enforcement is another matter. P4P incentives are especially challenging in the government and nonprofit sectors, which are homes for activities that pose more severe measurement and enforcement regulations. These result from the process of producers and consumers selecting the institutional forms with which they prefer to deal; private firms are most appealing when buyers and sellers are essentially equally informed about an output's quality. Buyers have incentives to select other providers, though, when they recognize their information handicaps *vis a vis* providers, and when they judge that government agencies or nonprofits act as useful protectors of under-informed consumers; there are inevitable uncertainties but, nonetheless, there is some degree of protection from gaming by better-informed, self-serving, profit-seeking, interests.

Differences in organization behavior, even within industries, can be real or illusory. The basic explanation is fundamentally simple – an efficient P4P reward system requires aligning incentives with advances on social goals. But that is not easy; the more costly it is to implement a reward system that *measures* performance well, even when that involves measures that are easy to *value* and that are largely immune from distortionary gaming that can be easily detected and prevented, facilitating P4P arrangements.

Preventing gaming is easier said than done. Only its degree of success, however, aligns high-powered rewards with advancement toward social goals. Otherwise, program incentives will fail to achieve their social objectives – or worse, will send the wrong incentive signals, misdirecting effort and undermining output quality, contrary to the rationale for particular incentives in the first place. If, for example, a nonprofit organization is subsidized to advance a specific public-interest goal, but the subsidy leads to increased gaming, the public interest may well be sacrificed.

Even today, these measurement and valuation issues remain prominent: How can and should we measure how "good" a hospital and any of its services are? A college and each of its many programs? A prison? Museum? Philanthropy? A Federal judge?

P4P Incentives in a World of Information Asymmetries (Inequalities)

Informational inequalities dog efforts to reward better performance. The 1970 Akerlof and Titmuss publications showed, in very different markets, countries, and social systems, how information inequalities between buyers and sellers can drive behavior in distinct directions.[8]

Akerlof examined, in his path-breaking work, "Informational Asymmetry and the Market for Lemons," how information and incentives interact and conflict. Focusing on the market for used cars in the U.S., he showed how a market can be weakened and even destroyed as a result of not just *imperfect* information on the quality of what is being offered for sale; is the car a "lemon," a "creampuff"? -- but the *inequality* of the information between prospective sellers and buyers, claimed Akerlof, typically favored sellers. Owners of cars that had been driven with great care and attention knew that they were selling cars of exceptional quality, but insofar as buyers were less informed, they would not pay more for a car of allegedly higher quality that they could not fully evaluate. So "cream puffs" would be withheld from the market, predicted Akerlof, as owners realized that their cars' values to *them* exceeded the market values of cars of similar age, model, mileage, and external appearance, that buyers could observe and value.

Across the Atlantic, and in the same year, 1970, Richard Titmuss was providing insights to the importance of information inequalities in a quite different market, for human blood, in the U.S. compared with the U.K. He explained the far higher prevalence of the hepatitis-B virus in the U.S., also highlighting informational asymmetries between buyers' and sellers' knowledge of the quality -- purity and safety -- of the blood being used for transfusions.

As in the used-car market, the culprit in the blood-quality market was seen as blood *suppliers.* In the 1960s the state of knowledge of how to prevent transmission of hepatitis-B through blood transfusions was rather rudimentary; the first hepatitis-B vaccine was not approved in the U.S. until 1981,[9] more than a decade after the Akerlof and Titmuss publications.

Titmuss had claimed that the basic hepatitis-B problem in the U.S. was the differential cost of information on blood quality available to buyers and sellers -- and the incentives for the better-informed party, generally the provider, to *not* reveal it. Rewards, he claimed, discouraged potential blood *donor*s from revealing their private information about the likely poor quality of their own blood — its contamination through the illicit effects of drug addiction and needle-sharing. U.S. donors, he argued, typically *sold* their blood, and had no incentive to disclose its likely tainted quality; they had reason to believe it was likely contaminated by the hepatitis-B virus, so it would be rejected if the truth about the donor's addictive behavior was revealed.

---

[8] Akerlof, "The Market for Lemons;" Titmuss, *Gift Relationship*.

[9] Beasley RP, "Development of Hepatitis-B Vaccine."

Not so in Britain. Titmuss emphasized that under its National Health System, *all* blood for transfusions was truly gifted -- supplied without recipient payment, current or future. Donors had no incentive to lie about their health, drug addictions, needle-sharing, or other unsafe habits.

Despite the similarities in Akerlof's and Titmuss' emphases on informational problems in undermining market efficiencies, and despite the fact that their works were published in the same year, 1970, it is not clear that they were even aware of each other's research, that both had focused on markets with significant informational *differences* between suppliers and demanders; both analysts saw the adverse incentive effects of the differences, and both identified the underlying problems of how to detect, measure, and value product *quality*. Moreover, when quality was better known by one stakeholder, but there was no incentive to reveal it to the less-informed party, the outcome would be an information chasm, retarding advancement toward the societal goals.

Both authors identified puzzles including: What do economics and social work have in common? And what do used cars and human blood have in common?

A great deal, it turns out. The commonalities were substantial, at least as of a half-century ago, when knowledge and technologies for identifying, measuring, and correcting product defects were far more limited than they are now -- whether in used cars, disease-tainted blood for transfusions, or other markets. But so were opportunities for someone with an information advantage to use it for private benefit – to develop and game an incentive system that rewards stakeholders who have information advantages over other market participants.

P4P is All About Measurements, Rewards and Incentives

Titmuss, from his base in the U.K., saw informational inequality among stakeholders as the underlying explanation for the greater prevalence of hepatitis-B in the U.S. than in the U.K. The key force was the disease's spread through blood transfusions in the U.S., of human blood that was contaminated with the hepatitis-B virus that led to fatal liver infections or liver cancers. But how?

Titmuss' explanation emphasized differences in incentives in the two countries and their effects on the safety of most human blood "donations." In the U.S. the donations of blood were essentially sales, not gifts, by blood *providers* who knew more about the safety of their own blood than did transfusion *recipients*. American donors, argued Titmuss, were generally aware of their needle-sharing practices, and other dangers of transmitting a disease carried by their blood. But they "needed" the money to finance their addictions. Moreover, at the time of his research, in the 1960s, the available tests for the presence of the hepatitis-B virus in a particular blood sample was of little value, for as I noted earlier, the shelf-life of blood after its withdrawal from a living person was shorter than the time needed to test its safety.

The blood supplied through the nationalized British National Health System was safer than what was being supplied by the American system, which was largely a product of decentralized private enterprise. There were potentially far-reaching implications of the institutional differences, many captured by the following question:

Performance Measurement, Valuation, and Gaming --Ownership Form Matters

Do information differences between consumers and producers of products of diverse ownership forms –- for-profit, government, private nonprofit –engage in differential gaming? If so, how?

The answers shed light on *The Perils of Pay-for-Performance: Why Strong Rewards for Performance in Government and Nonprofits Don't Work.* The hazards of expanding incentives used by profit-maximizing firms to incentivize employees of government agencies and nonprofits, are massive, in large part because of (a) the variety of collective goals they pursue, (b) the challenges of *measuring* and *valuing* each dimension of successful performance, and (c) the difficulty of allocating a supplier's overall success among the members of what amounts to a *teamwork* production process in which there is no single statistic that connects each input provider's participation to the overall success. Yet that aggregation is necessary if governments and nonprofits pursue ever-stronger links between their employees' individual productivity and the organization's performance goals.

So, does an organization's legal ownership determine its behavior, or vice versa? Contrary to common wisdom, it is not ownership form that determines an organization's behavior, but largely the converse, an organization's goals determine its choice of ownership form and the advantages and disadvantages each would bring taxes, subsidies, prohibitions, mandates, and other regulatory constraints, and employee motivations that affect the achievements of organization goals. These restrictions and opportunities accompanying each organization form differ, so a provider's selection of its organization form becomes a dynamic matching process between goal attainments and the choice of organization form through which to pursue them.

When a society's collective preferences support the position that consumers with the greatest willingness and ability to pay for particular commodities should win-out, those activities gravitate to the private enterprise sector, where organization behavior is least restricted -- top bidders win. But some trade is not simply an exchange between individuals pursuing their private goals, instead it involves provision of collective goods that are shared widely.

Some preferences, in short, are not simply for individual, private, goods or services that belong to a specific person, but are "collective," public," commodities that are shared, like clean air, global warming, international peace, internet access, and access to medical care that is generally viewed through a more egalitarian lens – judged to be desirably accessible to "all," not only to the wealthy. preferring a social system to which the rich and poor should have more equal access to, say, "high quality" medical care and higher education than would be achieved through unfettered private enterprise markets. Government and nonprofit providers are not restricted from pursuing greater profit, but they are encouraged by subsidies, tax exemptions, and regulations to provide services to buyers other than to those with the greatest willingness and ability to pay, and they do face taxation of profits from activities that are "not substantially related" to their tax-exempt 501(c)(3) status.

Yet the alignment of private rewards with social goals does not eliminate society's need to be leery of stakeholder gaming. In higher education, for example, admission standards might be undermined, and indeed have been, by admission officers who accepted bribes from wealthy parents seeking to "buy" admission for their children, or parents of applicants who have lied about their children's high school extracurricular accomplishments, to strengthen their case for college admission. Nearly 20 students at the nonprofit University of Southern California (USC), in Los Angeles, for example,

who began the 2019-2020 school year "not knowing whether they would be allowed to remain at the school or be expelled." [10] Improper, even illegal, gaming of admission practices had been claimed.

Gaming and Incentives to Undermine Other Incentives

Gaming, based on asymmetric information among potential buyers and sellers, limits trade. Yet it is widespread, not limited to any particular industries, organization forms, countries, or products. Laws, regulations, and penalties do not ensure that performance-based incentives will be effective, nor that efforts to circumvent them will be, effective, regardless of organization form or the countries involved. To see why, consider this example:

In 2016, the United States Congressional Research Service reported on intentional and illegal gaming by a major international private car-manufacturing firm, the Volkswagen (VW) Automotive Group. VW admitted to having installed special computer software on several of the types of diesel-fueled engines it was testing; the software was designed to act as a "defeat device;" the purpose being to undermine the social goals that the U.S. regulator, the Environmental Protection Agency (EPA), was pursuing -- to cut air pollution and to show that the VW diesels were increasing travel efficiency, measured by miles per gallon of diesel fuel used.

VW was playing multiple roles; as a producer and seller of cars it stood to benefit from evidence of its cars' fuel efficiency, As a principal stakeholder in the outcomes of the tests. VW had knowingly and illegally misled two other sets of stakeholders – potential car buyers, who favored greater fuel economy, and the EPA, which favored less air pollution.

The various stakeholders were engaged in a struggle over the distribution of benefits and costs of selling more VW cars, cutting the air pollution caused by the increased sales, and potential VW car buyers, and the federal EPA, who valued better fuel economy and cleaner air. By exaggerating to prospective car buyers, the number of miles per gallon of diesel fuel a VW owner could expect, and by *understating* in their reports to the EPA the air pollution that a VW diesel engine was generating, VW was benefitting financially through each of three routes:

The VW *defeat-device* software did "work," at least temporarily. It detected when a car engine was actually being tested for regulatory compliance; then it activated pollution control devices that reduced the *measured* tailpipe emissions (of nitrogen oxide) in some 600,000 diesel vehicles. VW admitted to gaming[11] and to what came to be termed the "biggest [fraud] in automotive history." [12]

Gaming—Legal or Illegal?

The VW pollution gaming was illegal, but not all gaming is. And it worked for VW, until a whistleblower within the company released information on its manipulations of the pollution-testing process to deliberately understate the engine emissions reported to the EPA and to prospective VW car buyers -- which *over*stated the fuel mileage it reported, and *under*stated the cars' air pollution

[10] Taylor, "Parents Try Paying to Avoid Piper," and Taylor, Admissions Scandal Looms"
[11] Canis et al, "Volkswagen, Defeat Devices"
[12] Ewing, "Volkswagen's Swift Journey"

emissions. As of March 2020, "Volkswagen AG said its diesel cheating scandal had cost it 31.3 billion euros ($34.69 billion) in fines, penalties, financial settlements and buyback costs." [13]

Five years later, as the year 2025 approaches, the Volkswagen case has not disappeared. The company's former chief executive, Martin Winterkorn, went on trial, on September 3, 2024, facing "criminal charges including fraud, market manipulation and making false statements when, in 2014, he became aware of software designed to illegally cloak emissions that exceeded limits imposed by European and U.S. regulators ….The court has scheduled about 90 days of testimony … which is expected to last until September 2025." (Add footnote 13a here, as follows:

But gaming is not limited to private firms or their interactions with government regulatory agencies, consumers, or even the world population concerned with climate change. All had incentives to develop information superiorities and then take advantage of them, as was the case with *nonprofit hospitals* in Illinois and their incentives to exaggerate their provision of "charity care" to gain rewards through property tax reductions. And we will soon see in another paper in this series, the gaming by public school teachers and administrators who would also be rewarded if they cheated in their grading by raising student scores on their tests to "show" that the schools were performing better.

Organization Conversions: Does Ownership Form Matter?

Nonprofit and government organizations dominate many industries: hospitals, k-12 schools, colleges, prisons, museums, symphony orchestras, libraries, courts, and more. Each has advantages and disadvantages compared with each other and with private firms, when they coexist in the same industry and market. Each also has the legal freedom to convert from one ownership form to another, though with restrictions, impediments, and tax consequences.

A change in ownership form is a change in constraints and incentives. In many conversions from one institutional form to another, the differences affect decisions on which specific services to provide and how to value and report them. So if the management of an organization sees untapped opportunities from a conversion, it can change its legal form, though not costlely. Conversions of *hospitals* from nonprofits to for-profits have been particularly common, though surprising, in light of the loss of tax exemptions from profits on activities that were "substantially related" to their tax-exempt mission, and the lost donations from private donors who itemize their tax-deductible donations to "charities. "

The *Perils of Pay-for-Performance*, especially for government and nonprofit organizations, are, in short, not the result of inept management. They are the result of differential incentives that induce a sorting process that matches organization forms with managerial opportunities and incentives. Gaming is encouraged in the process -- some intended by the legislative framers of the rules for conversions, rewards, and penalties, some not; some gaming is lawful, some is not. All are the outcomes of a Pay-for-Performance matching process within a dynamic economy that shapes opportunities and incentives to choose among them.

In the years prior to the Akerlof and Titmuss publications in 1970, information asymmetries between suppliers and demanders had received scant attention. So, too, had the logic of producers' and

[13] Business Insurance, "Volkswagen Emission Scandal".

consumers' choosing among organization forms with different constraints and incentives. To be sure, though, consumer information differed in the evidence supporting sellers' claims for their products' safety, efficacy, and quality, as technological advances and measurement methods and accuracy improved, creating and destroying information asymmetries between sellers and buyers.

But long before this research, indeed by 1906, there was already growing recognition that in what we now call the pharmaceutical industry, sellers of "medications" were making outlandish claims for their product quality -- safety and efficacy – though with essentially no support from research, clinical trials, or other evidence. Information inequalities, however, were undermining consumer confidence in sellers' advertising claims for what had come to be called "snake oil," stoking the demand for government regulation of medicines and food products.

Medications, Medical Care, and the Food and Drug Administration

The 1906 passage of the Pure Food and Drug Act replaced the free entry and unrestricted claims by sellers that their products safely cured all sorts of illnesses and imperfections, even when there was essentially no supporting scientific evidence. No longer could a Sears Roebuck & Company mail-order catalogue assert that one of its medicinal products brought miraculous results against multiple ailments, as it did in its 1902 volume touting "Dr. Rose's French Arsenic Complexion Wafers," which it claimed was "perfectly harmless," possessing

> *" …the **Wizar**d's **Touch** in producing, preserving and enhancing beauty of form and person in male and female by surely developing a transparency of complexion, shapely contour of form, brilliant eyes, soft and smooth skin, where by nature the reverse exists. … their effect is simply magical…. Ladies, … you can make yourself as handsome as any lady in the land by the use of our French Arsenic Wafers*.[14]

There was not even a hint of the danger of ingesting arsenic.

The National Institute of Health (NIH) had been established earlier, in 1887, with an impressive name but with only a one-room office and little authority. The U.S. Food and Drug Administration (FDA) followed nearly 20 years later, in 1906, gradually becoming the principal government watchdog over the development, production, and distribution of medications that had been shown to be "safe" and "efficacious" in preventing or mitigating specific illnesses. From its humble beginning the NIH grew to 18,648 employees in 2013 and subsequently to over 22,000, not counting academic grantees who were using NIH research grants to advance health care knowledge and applications. The Covid-19 pandemic of 2020-2022 further increased public attention to the roles of the FDA and other government regulatory agencies in approving vaccines against viral diseases, including development, along with the United States Center for Disease Control and Prevention (CDC), of recommendations that the public wear face masks, be vaccinated and "boosted," and maintain a 6-foot "social distancing."

Much has changed since the industrial revolution of the late 18th and 19th centuries, in technological knowledge about the prevention and treatment of diseases as well as other elements of

---

[13] Ewing, Jack, and Tatiana Firsova. "Former Chief of VW Faces Fraud Trial Over Testing," New York Times, September 4, 2024, pages B1, B5.

[14] Sears, Roebuck and Co., Catalogue, 1902 Edition.

science and technology. Improved instruments have been developed for *measuring* medical product performance, world trade has expanded, facilitated by advances in navigation measures and techniques, as well as in power sources replacing wind-powered sailing vessels with steam-powered ships. And education has been increasing along with life expectancies -- all making future job-market patterns less predictable. The historic stability of pre-industrialization labor markets has been replaced by volatile employment opportunities as workers live longer, work longer, and are rewarded for their resiliency in converting schooling obtained in their youth into new, more-modern, skills. Increased education became the favored road to economic opportunity.

Education and its Expansion in a Changing World

At the beginning of this century, in 2001, the acronym, STEM, (Science, Technology, Engineering, and Mathematics.) was introduced by the U.S. National Science Foundation (NSF) to capture the directions the "experts" foresaw as re-shaping the future labor force and employment opportunities in the U.S. and elsewhere in the industrial world.[15]

As governments at the federal, state, and local levels were becoming increasingly prominent in the industrializing and science-based economy, so have new forms of organizations evolved. In particular, *private nonprofits* emerged as hybrids combining the decentralized flexibility and agility of private firms with the power and authority of collectively owned government units.

In the U.S. the *education* sector has expanded in multiple directions and organization forms, pursuing broadening goals. In 1940, shortly before the U.S. involvement in WWII, fewer than five percent of all adults aged 25 and older had graduated from college – 5.5 percent of males, 3.8 percent of females. By 2021, college graduation had become commonplace, no longer just an accomplishment for the elite, but soaring to 36.6 percent for men and 39.1 percent for women, some six to seven times the 1940 levels. And now, the college graduation rate has become greater for women.[16]

At the post-high school level more new educational options emerged. Two-year "community colleges," sometimes termed "junior colleges" and "trade schools," had emerged decades earlier, in 1901, in Illinois. But it was not until the aftermath of WWII and the first of a series of *G.I. Bills of Rights* became national law that financial access to two-year and four-year college education for millions of honorably discharged military veterans dramatically opened new educational pathways beyond high schools.

In the 1960s there was an explosion of the numbers of community colleges and their enrollments. Two-year colleges were opening at the rate of one per week.[17] The decade of the 1970s saw their enrollments grow from 1.6 million students to more than 4.5 million (Brint and Karabel, 1989). These two-year, largely tuition-free, community colleges had been predominantly vocational institutions teaching specific job skills, but between 2010 and 2019 their enrollments declined by more than 1.6 million students, as the belief increased that future economic change would favor employment opportunities for youth with 4-year baccalaureate degrees.

---

[15] Hallinen, "STEM education curricula."
[16] U.S. Census Bureau, Education Attainment Tables, 2021
[17] Cohen, *The American Community College.*

In 1988 the Report of the Commission on the Future of Community Colleges (CCs) defined those schools not only as geographical locations but also as a climate for post-high school learning. CCs in America were enrolling more than 10 million students annually – 44 percent of all college undergraduates and 50 percent of the incoming freshmen. (Brint,1989). By 2020 there were 942 public CCs in the U.S., offering a more accessible, affordable, and less time-consuming path to a two-year certificate program, although as of 2019 fewer than 40 percent of their students were earning even a certificate, let alone transferring to a four-year college and receiving a bachelor's degree within six years of CC enrollment.

These expansions have taken place through a wide range of organization forms, including, in addition to public CCs, state-owned universities and their Federally subsidized "land-grant" universities, and private *for-profit* schools such as the University of Phoenix and Capella University, and private *nonprofit* schools such as Southern New Hampshire University, with its 2020-2021 enrollments of over 41.000 full-time students and over 93,000 part-time, as well as over 1,500 other nonprofit colleges, in addition to trade schools teaching specialized skills for such jobs as barbers, beauticians, and construction workers.

Beginning in 1944, with the end of WWII in sight, the roles of education entered a historically unprecedented era; millions of veterans were made eligible for Federal grants to finance completion of high school as well as higher education, in addition to covering housing and health care, under the first *G.I. Bill of Rights.* Later legislation extended similar G. I. Bills to veterans of the Korean and Viet Nam Wars, to their families, and then to victims of the "9-11" (2001) bombings of the two New York World Trade Center Towers and related targets including the Pentagon, in Arlington, Virginia. Benefits were also being extended beyond the law's "key provisions: education and training, loan guaranty [sic] for homes, farms, or businesses, and unemployment pay." [18]

Higher education was at the core of national planning for the uncertainties of a post-WWII economy, increasingly being seen as the road to economic growth but also to avoidance of a repetition of the Great Depression of the 1930s. Education was the arena in which the interplay of uncertain peacetime job market *demand* and the predictably massive increase in civilian labor force *supply* would play out, as millions of largely young and unskilled former military personnel would be discharged and enter the civilian labor force. A severe post-WWII depression was regarded as likely -- unless the government intervened massively. Which it did

There was *no* post-war depression. That was at least partly because of the unprecedented civilian educational opportunities offered to the honorably discharged military personnel through the G.I. Bill of Rights. It was financed by the Federal Government and accepted by about half of the eligible former service personnel were mostly men, for women had not been subject to the military draft, nor were they accepted as volunteers for combat.

The G.I. Bill brought two important effects: First, the former troops who returned to a school as "students" were not counted as being in the Labor Force, and so were not counted (measured) as either "employed" or "unemployed;" the nation's *reported* unemployment rate was thereby held in check

[18] U.S. Department of Veterans Affairs, *Born of Controversy: The GI Bill of Rights.*

compared with a much-feared alternative in which there could have been vast numbers of discharged service personnel who were *unemployed.*

Second, subsidizing higher education for veterans brought social and economic benefits from the nation's investments in future labor force productivity growth, increasingly being promoted as "human capital." Education was being viewed as a means of increasing labor-force skills and productivity, thereby reducing the probability of a return to the Great Depression of the pre-WWII 1930s.

Post WWII Education and Medical Care

After WWII, higher education entered a new phase. Medical care also came to front stage as the national system of Veterans Administration (VA) hospitals became the Department of Veterans Affairs, and expanded to treat the millions of former military veterans who were being honorably discharged and with increasing need for medical care, not necessarily related to their military service. The hospitals came under increased pressure as veterans sought not only more schooling but more medical attention and promptly. Waiting times to see a physician soared; average wait times to be seen by a primary care physician reaching 115 days, nearly four months, and bringing growing political demand for more responsive attention.

But how? Administrators of the Phoenix, Arizona Veterans Administration (VA) hospital responded, though not in the intended form. Faced with a budget constraint, the hospital developed an alternative it gamed the record-keeping system and its related rewards, to improve performance— by cheating .The VA Hospital intentionally under-reported patients' actual waiting times. Here was the way:

The hospital was keeping two waiting lists; one was a *secret waiting list* of patients' *true* waiting times, the other reported shorter, but false, waiting times to the central VA system office. The hospital was hiding an embarrassing truth; its reports were understating actual waiting periods, to bring rewards for "better performance."

Decades passed. In 2014, then-President Obama became involved, vowing to determine whether there was intentional cheating in the form of systematic under-reporting of wait times in official reports by the VA hospitals. There was.

Local VA hospital administrators were taking advantage of their informational superiority to game hospital record-keeping; they had greater knowledge of the decision-making and reporting systems they were using, compared with the rules established by the VA System in Washington, DC. The reports on patients' appointment waiting times were being falsified to show that the waiting times to meet with a VA physician were decreasing and hospital performance was improving; but the real goal was to shorten *actual* waiting times, not simply hospitals' *self-reported* and downward- biased wait-time reports.

Another goal and dimension of hospital performance was to *cut the costs* of the quicker care. That was achieved. So there was an administrative "victory" of sorts, at least for the VA hospital administrators; by under-reporting the true wait-times, the VA Hospital did bring private bonuses to its administrators for their alleged successes. But the VA system goal in instituting bonuses was not to hand-out money to administrators; it was to expedite veterans' medical care while controlling costs. Bonuses were incentives, not goals.

Whatever industry may be pursuing greater productivity and lower costs, and regardless of how performance is rewarded, the challenges are essentially the same: achieve the social goals while suppressing gaming. Depending on the industry involved, a stakeholder might be, for example, an

official of a VA hospital, a local police department official whose performance is measured and rewarded by the reduced "crime" rate in the district, a college president whose compensation is linked to the school's increased "graduation" rate or its increased national ranking, or a nonprofit charity's performance might be judged by how small a fraction of its total revenue is spent on "administrative costs" – providing new opportunities for gaming. But whether decisionmakers are government agencies, nonprofit organizations, or private for-profit firms, and whether the gaming involves violations of laws or of internal VA rules, stronger incentives invariably encourage self-serving actions.

P4P, Conflicting Goals, and Heisenberg's Uncertainty Principle in Quantum Physics

The Heisenberg Uncertainty Principle and its implications for simultaneous *measurements* of variables that are interdependent, each affecting the other, is more than perilous. He convinced the atomic physics community that those simultaneous measurements are *impossible;* the view still holds.

In today's context, using P4P incentives to optimize efficiency turns out to highlight other goal interdependencies. Each goal involves measurements and valuations of a particular resource's contribution to that goal. Increasing measured performance in one dimension can be expected to cause a decrease in another, and the range of effects can be enormous.

My conclusion is *not* that the search for better ways to measure, value, and reward performance should be abandoned because of their colliding interdependencies. Rather, public policy should be sensitive to a troubling dilemma: there are unavoidable side-effects of *over*-rewarding advancement of goals that are easily measured, while *under*-rewarding advancement of goals that are costly to measure, , and so are often overlooked.

College education is a case in point: Measuring the number of students in a college class is easy, as is the cost-saving from larger classes; but measuring the *effects* of larger classes on student "learning" is not. So while there is an *easy* path to *cost cutting* -- *enlarge class sizes* -- that disregards the likely negative effects on educational quality, of the decreased interactions of students and faculty in larger classes. Those effects are costly to demonstrate, let alone to value, but that does not imply that this form of performance is worthless; measuring, valuing, and rewarding performance in all its forms is a series of gameable hurdles, their far-reaching effects are visible throughout the economy:

- in *k-12 schools*, student scores on standardized tests have been altered surreptitiously by teachers and other school staff, in Atlanta, to demonstrate improved school performance;
- in *hospital*s, where provision of charity care was the sole measure of performance for determining exemption from real estate taxation in Illinois nonprofit hospitals;
- in *police departments*, where the numbers of reported crimes and arrests have been used to gauge police success in controlling crime while avoiding racial and ethnic profiling, as in the case of the Chicago Police Department gaming of its reports; and…
- in *museums*, where "attendance" and "donations" are often used to measure "success."

The opportunities and incentives to game a reward system for particular stakeholders are threads that connect industries and organization, including, for example:

a) in *k-12 schools*, where teachers and administrators have reaped rewards for their accomplishments, as measured by improved student scores on standardized tests, even

when the scores were achieved by teachers who knowingly cheated, erasing students' wrong answers and substituting correct ones:

b) in *colleges,* where higher student Grade Point Averages (GPAs) were interpreted as measures of increased student "learning" rather than as simply grade inflation;

c) in *policing,* where a reduction in the number of "thefts" was used to show greater citizen safety, by convincing a community resident that her reported "theft" of her purse was really only a "mysterious disappearance," which is not a "crime";

d) in an internationally renowned *museum* that substantially increased its admission price, it then studied the effect on attendance, finding, to its surprise, that only a very specific attendee group had been affected – those people who wanted to use the restrooms.

Other papers in this series provide a deeper dive into the scope of P4P gaming and affected stakeholders. The fundamental challenge is not the abstract possibility of mis-measuring and mis-rewarding performance; it is the *costs* of developing and implementing measures that capture true, unbiased, performance measures. Easier said than done.

I will revisit these measurement issues, emphasizing the different advantages, disadvantages, and gaming opportunities they reflect, for alternative organization forms -- private for-profit, government, and private nonprofit. They have different legal restrictions, including the "non-distribution" constraint (NDC), which prohibits *nonprofits*, but not private *for-profit* firms, from pursuing profit and distributing some or all of it to their managers, trustees, or others who control the organization. To be sure, of course, the costs of enforcing the NDC and other restrictions are often substantial, and they create new opportunities for gaming.

Sources of P4P "Perils" -- Inaccurate and Incomplete Measurements

Some forms of performance are easier to measure and value than are others. Those that are virtually costless to measure, value, and convert into rewards (or penalties) for *better* (or worse) performance, are quickly adopted. Toward the other extreme are performance elements that are so costly to measure, value, and reward accurately that they bring incentives to evade or even disregard them.

Therein lie the perils of a P4P system that carries strong incentives for forms of performance that may be "better" by some particular standard but are costly to achieve without cheating or otherwise exaggerating accomplishments.

Measurements that are inaccurate or incomplete are the culprits. They divert resources *away from* activities that are more costly to measure and reward or punish accurately, and *to* activities that are easily measured and rewarded. Teachers in elementary schools, for example, have been incentivized to have their students excel in math and science, for which there are standardized tests that are used to reward *teachers'* "excellence," while devoting little attention to the arts and humanities, where there is little or no consensus on how to gauge student and teacher performance. In museums, performance quality is typically treated as a matter of peer opinions; there are no standardized tests of "learning;" in the public-policy arena, government "performance" is often identified in such amorphous terms as anti-racism, greater income equality, and world peace.

True, each of us may feel confident that we know how to measure and value the dimensions of performance quality; yet there may be little consensus on how to obtain the necessary information

while avoiding gaming or asking administrators to evaluate their own performance or that of friends or colleagues; however, these are clearly dubious bases for rewarding performance. They establish incentives for gaming performance measurements, including stakeholder creativity in developing new ways for stakeholders to advance their own interests.

In short, there are reasons to doubt the conventional wisdom that using stronger performance-based incentives assures more efficient production and higher quality: (a) employees may not know the system that determines their compensation, which includes how the multiple dimensions of output quality are measured; but (b) even if they do know, they typically have incentives to respond strategically, pursuing their own self-interests rather than those of the employer or regulator, and (c) incentives can be gamed in many ways, including, as noted earlier, by changing ownership form, regulations, and their enforcement.

Information Asymmetry and Gaming of Rewards: Does Ownership Form Really Matter?

Do stronger rewards for better performance differ among for-profit firms, governments, and private nonprofits? They do*,* but not for the commonly held reason of differential efficiency.

Private firms are not inherently more efficient; in a modern economy there is a dynamic process of organization innovation leading to differential expansions and contractions of for-profit, nonprofit, and government roles. Central to the realignment process are measurements of performance, its value, and the translations into producers' incentives. Effectively rewarding performance in a changing economy requires adaptations to the dynamics of choice among ownership forms. Government regulation of automobile engine air pollution, for example, has replaced the unregulated air pollution market amidst worldwide concern with global warming.

When quality of an output is multi-dimensional, when the attributes are interdependent, and when there is little consensus on how to measure and value each element of performance, consumers are typically under-informed relative to producers, about appropriate measurements, as Akerlof emphasized. Consumers are thus vulnerable to having their information handicaps used against them under a P4P incentive system. Bear in mind, though, that the goal of rewarding better performance is to guide resources to their most valuable uses, not to generate private profits. But just as the social goals for lighthouses and beacons are to guide ships or aircraft through risky routes, mismeasurements *assure* misguidance.

An efficient reward system thus requires that guidance incentives be aligned with the values of *all* dimensions of performance*,* not just *some.* Otherwise, producers will respond to the more highly rewarded elements that are also more easily achieved, systematically overlooking the others. And they do, in ways that are centuries old.

Measurement *cheating,* by innkeepers, was portrayed in a Polish Church painting in 1699. The depiction was of intentional overcharging by an innkeeper who "never poured [a] full measure" of vodka into a glass. An inscription on the painting read "She never poured full measure." The vodka drops held

back by the innkeeper were more than a financial transfer from the customer to the innkeeper, but also a form of tax that reduced caloric intake by the "undernourished peasant." [19]

Gaming can and still does take the forms of manipulating both the comprehensiveness and measurement accuracy of output quality as well as quantity. Consider the opportunities open to a meatpacking firm that packages ground beef for retail grocers: a consumer might expect it to be in a package labeled with the weight -- say, 3.0 pounds. But there are uncertainties: what is the *exact amount* you actually got? Would you carry your own scale, and if the weights shown differed between your and the seller's, which would you believe? Does it really matter?

What can be said about the "quality" of a product, however, reflects the fact that, as in the ground beef case, the product has more than a single, homogeneous component; it is a mixture of beef and fat, and fat is cheaper than beef per pound. So, a profit- maximizing meatpacker has a financial incentive to game its information advantage over consumers by decreasing the ratio of beef to fat. Over time, better measurement techniques have evolved, as has stronger government regulation of food content, and while consumer learning may narrow the information gap, that may be a slow, costly, and unreliable process; it was in the case of the European innkeepers' short-changing patrons by chiseling on the quantity of vodka in a "serving."

Even now, the *price* of a "unit" of a commodity, while seemingly easy to observe, can be quite misleading; Whether in a grocery market, a college, hospital, jail, or any other industry, sellers have the incentive to over-report the content and quality of the more costly components, e.g., "ground beef" in the grocery market, and under-report the fraction of less-costly ingredients such as fat, to hold down the aggregate price per unit of weight., This gaming process reflects the common informational advantage of sellers over buyers, as Akerlof emphasized. Even when a third-party regulator or information-provider such as a government consumer protection agency or *Consumer Reports* Magazine, is monitoring the information-provision process, incentives favor sellers; they typically sell small quantities of their product to each of a large number of consumers, none of whom has a major incentive to detect and prevent chiseling or cheating.

One of the variables used most commonly to convey ostensibly helpful information between sellers and buyers is, not surprisingly, its "price." The principal attraction is its *apparent* clarity; what must I pay to purchase a specific item? But its *quality* can be gamed; in higher education, for example, an increased number of students in a class is relatively easy to observe, but its effects on various types of "learning" are not, comparably to the relative ease of determining the total weight of a package of "ground beef" but not the component weights of beef and fat.

The industries I examine, and their performance measures, are quite different in some identifiable dimensions, very much alike in others. They differ in their use of strong and weak incentives to reward "better" performance. They range widely in their ease of gaming the performance-based rewards, as well as in their reliance on high- powered incentives to encourage better performances. Why, with what results, and with what future?

---

[19] Kula, *Measures and Men.*

Distorted measurements and gaming are basic ingredients for an *in*efficient incentive system: information about productive performance is commonly *costly* to measure accurately, without systematic *bias,* because the stakeholders who are best informed about true effort and outcomes often do not have the incentives to reveal them. Rather, they have the incentive to overstate their favorable accomplishments that generate higher rewards, and understate the performance shortcomings that would diminish rewards. In the process, forms of performance that are costly to measure and value get little attention.

So trying to encourage greater efficiency through stronger incentives often accomplishes the opposite, even apart from outright cheating. That was the saga of WWII veterans who, after their discharge at the war's end in 1945, were waiting long periods, often months, to see a physician at a VA Hospital, while at the same time the hospital was *falsely* reporting their successes in *cutting* waiting times; more about this later, in the case of the Phoenix, Arizona VA Hospital, although at another hospital, the Hines VA Hospital in suburban Chicago, a U.S. Senator from Illinois, Mark Kirk, reported that boxes of VA electrocardiograms (EKGs) had been taken but not read, and several veterans had died long after their EKGs "had been taken but not read."[20]

Similarly, in *k-12 schools*, children were reported as performing better on standardized tests than they had previously, as teachers and administrators improperly and in some cases illegally, corrected student errors to improve their test scores; and in state prisons operating under contracts with *private for-profit* prison companies, contractors recognized that under their contracts with state Bureaus of Prisons, the firms could increase their profitability by treating ill inmates inside the prison rather than transferring the prisoners to a hospital for treatment to be paid by the prison contractor.

Rewarding "Performance" -- Perils and Attractions

Now back to the Provena case to delve further into the problem of ambiguities. In that case, Provena Medical Center's *performance* and that of other nonprofit hospitals in Illinois, in terms of its services to the poor and uninsured, did not justify the state's exemptions of the hospitals from property taxation. The State of Illinois Department of Revenue used hospitals' own reports to conclude that a number of nonprofit hospitals were not providing "enough charity care" to warrant their being relieved from real estate property taxation. But what was "enough?" And how was a nonprofit hospital supposed to determine the monetary value of its actual charity care? And how were hospital reports supposed to be *monitored and audited*?

Provena appealed the state's negative Revenue Department decision on property tax exemption; it lost again.[21] The legality of linking a hospital's *self-measured* performance to the dollar amount of its charity care and to its financial reward in the form of property tax exemption, was not only upheld but would apply to other nonprofit hospitals. Showing how the link produced unintended and inefficient incentives not only for other hospitals but for other nonprofits aiding the poor, underscores the perils of Pay for Performance (P4P) in social services. Adoption of such an incentive to meet or exceed some specified level of charity care to qualify for property tax exemption or some other government subsidy, was adopted intentionally to incentivize hospitals, in this case, only nonprofits, to at least meet specific

[20] Smith, "Sen. Kirk calls for Hines VA Hospital director to step down.
[21] Supreme Court of the State of Illinois, "Provena Covenant Medical Center"

public service levels, and presumably in particular forms, and to do so while presumably avoiding providers' incentives to game the performance measurement process.

Incentives to Game a Reward System

In the Provena Hospital case it was easy to see the strong incentive for many nonprofit hospitals to search for a low-cost way to improve their reported "charity-care" performance to regain, or at least retain, the property-tax exemptions they were in danger of losing. Alternative cost-accounting procedures might help, perhaps by providing acceptable allocations of a hospital's overhead costs between patients who can and cannot pay the hospital's prices. When another accounting procedure permits recording more joint, overhead, costs of health services to the uninsured poor, which could result in retention of a hospital's property tax exemption.

Seven months before the Illinois Supreme Court decision in the Provena case, three other hospitals had lost their property tax exemptions, all for the same reason, "insufficient charity care." One, Prentice Women's Hospital, a subsidiary of the Northwestern Memorial Hospital system and located on prime land in downtown Chicago, was hit with a $66 million property tax bill for tax years 2008-2011.[22]

Even if it were entirely clear what medical services a hospital actually provided to poor, uninsured, patients, and even if there were no problems of determining when a patient really is uninsured and poor – big "ifs" – measuring charity-care performance would face yet another hurdle: how to measure the "value" of the services provided? A gigantic, ambiguous, task.

Nonetheless, while the Illinois Supreme Court decision had required that the *value* of a nonprofit hospital's total provision of charity care, somehow gauged, must exceed its property tax bill, to justify exemption from the tax, the state legislature and the state Supreme Court were silent on how the value of the care should be calculated. This allowed hospitals to use self-serving discretion in their valuations. That task was apparently left to hospital accountants, the implicit assumptions being that the valuation procedures were unambiguous and "scientifically" determined, and that there was no important discretion available to a hospital to game their performance reports, no inflating the value of charity care to at least meet the threshold level for property-tax exemption. But those assumptions were wrong; gaming was more than a theoretic option.

The Illinois decision to gauge a hospital's performance by its provision of *free* services to the poor was narrowly defined but ambiguous in practice. It was narrow in that there were other socially collective goods that might have been included to justify property tax or other exemptions. Indeed, under Federal law on tax exemption of nonprofit organizations such as hospitals, universities, and other "charities" that are tax exempt under section 501(c)(3) of the U.S. Internal Revenue Code, a nonprofit could qualify for tax exemption by providing "community benefits," not necessarily only "charity care." Moreover, there was and is no standard for how those benefits *should* be defined, measured, and valued to determine the hospital's total social contribution and whether it was, under the Illinois law, at least equal to the hospital's property tax bill, to cancel it. Otherwise, the full property tax bill was due, no exemption.

---

[22] Carlson, "Setting a standard."

In short, the *good performance* required to justify property tax exemption meant only one thing – provide *free* care to the uninsured poor that was "worth" more than the property tax bill.[23] The value of care required for that exemption was also seemingly unambiguous -- whether the hospital's "expenditures" on charity care was at least equal to the property tax revenue that the local government would lose through its tax exemption. No other form of performance counted. And if a hospital lost its tax exemption under State law, there was only one way to regain it; provide *more* charity care. Or so it seemed!

True, good performance of a nonprofit hospital was defined by the State with historically unprecedented precision. What was far from evident, though, was exactly how each hospital would respond to the incentive to reach its tax-exempt threshold -- directly or through gaming. There was reason to be skeptical, for there were gaming options:

- *Option 1: Pay the tax bill and move on*—treat the property tax bill as a "fine" for providing too little charity care, but do nothing to change future services to the poor. The extra expense might be financed by cutting expenditures on other "unprofitable" programs such as community health education, if they exist.
- *Option 2*: *Admit more charity-care patients or treat the poor and uninsured more "fully," to reach the expenditure threshold and restore the hospital's full property tax exemption.* But measuring how much a hospital actually increases the *real* dollar value of its charity care is no simple or gaming-proof task.
- *Option 3. Gaming the calculation of "performing well."* If a hospital reports to the state that it "spent," say, $1.65 million, on charity care in a given year, what exactly does that number mean, how was it calculated, and how *should* it have been calculated? The answers are susceptible to manipulations that can advance a hospital's self-interest in retaining its property tax exemption.
- *Option 4. Lobby the state legislature to reverse its prior decision and permit granting of property-tax exemptions in return for a nonprofit hospital providing a wider range of community services.*

An example: consider the opportunities a hospital had, under the Provena decision, to game its P4P reward structure by using an accelerated accounting method to increase the reported depreciation expenses of its buildings and equipment, in part to report more dollars' worth of charity care. A number of alternatives were and still are legally available, although there are explicit limits.

It is comforting to think that everything a hospital spends money on is either charity care or not, but reality is more complex. When a patient does not pay the full hospital bill, either directly or through insurance, someone must determine how to record and report the expenses. The two principal candidates are "charity care" and "bad debt." If there are differential rewards for the choices, we can expect different decisions and rewards.

---

[23] Frost, "Legislation Defines Charity Care for Hospitals"

Until the legislation underlying the Provena case was repealed, a nonprofit hospital's annual property tax bill was tied to its reported "expenditures" on charity care, *not* to its *uncollected* patient bills (bad debts). But the distinction was frequently not clear, and nonprofit hospitals had the financial incentive to justify using a combination of the two measures to meet the charity-care hurdle and qualify for property-tax exemption.

Quid-pro-Quo Subsidies by Governments—an Untapped Revenue Source

The impact of the principle underlying the Provena decision was potentially immense – grant a nonprofit organization, not necessarily only a hospital, exemption from some tax – not necessarily only on real estate -- *if doing so* passes a benefit-cost test showing that the State's loss of tax revenue was exceeded by its gain from shifting the fiscal burden to another source.

The principle of demanding a *quid pro quo* from any subsidized organization, not only a hospital, for it to receive exemption from some tax had and still has widespread potential applicability. It could be applied to any industry, any tax, any state. Well before various fiscal attacks on hospitals, tax-hungry legislatures in Massachusetts as well as the federal government were moving toward unifying the tax-system benefits they offer in health care with another industry, higher education. In that industry, better performance had come to emphasize increasing the numbers of incoming students, graduates, and their subsequent lifetime earnings. These came to be associated with performance measures such as a college's 6-year graduation rate, and more college "graduates" receiving bachelor's "degrees" rather than "certificates" of lesser accomplishment, such as from a Community College or trade school.

In 2008 the U.S. Senate Finance Committee set out such a plan for higher education; it would tap the assets of "wealthy" colleges' performance by incentivizing them to spend-down their "endowments" in order to cut tuition and increase student financial access. The precise measure of a school's wealth and how it could be taxed were matters that would attract substantial attention in the next decade. It eventually led to the Tax Cuts and Jobs Act of 2017, which imposed an excise tax of 1.4 percent on the investment income, not the assets, of schools with at least 500 students and endowments worth at least $500,000 per student.[24]

Yet the fundamental concern that has continued to this day is the seeming paradox that even the wealthiest of colleges have been raising tuition even while adding to the schools' endowment wealth and reaping rewards from government through nonprofits' exemption from taxation of corporate profits from activities that are "substantially related" to their nonprofit, tax exempt, purposes, their capital gains, or their real estate. [25]

In effect, links were being considered between a college's *performance* in its tuition and student-aid policies, and its *rewards* from local, state, and Federal tax systems. It was also becoming increasingly evident that nothing prevented a government from expanding its P4P links between measurements of colleges' educational "performance" and the overall tax system.

But why only in hospitals and higher education? Why not extend the principle to reward a much wider range of performance measures and economic sectors -- to nonprofit nursing homes, museums,

---

[24] U.S. Internal Revenue Service, "Excise Tax on Net Investment Income of Private Colleges"
[25] Weisbrod, Ballou, and Asch, *Mission and Money: Understanding the University*.

symphony orchestras, and other nonprofit and government activities? The tax system has vast potential for adopting a more wide-ranging *quid pro quo* policy: either improve your performance, as we, the state, define and measure it, or be subject to some additional tax.

This can work, but at a cost; it inevitably encourages socially inefficient gaming and self-serving accounting practices – and not only in hospitals.

The Property-Tax Threshold: a Form of "Pass-Fail" Test Score

The Illinois law amounted to offering a reward for a "pass" score on a test; improving performance – reporting more charity care -- mattered only if it led to crossing the pass-fail threshold. Only then would the hospital's property-tax bill be affected. Improving a still-*failing* grade would not be rewarded with an additional tax exemption; neither would improve on an already-*passing* grade. Incentives would not encourage provision of more charity care unless the hospital (or, in another industry such as a college, museum, or court) was close enough to the tax-exemption threshold to make crossing it feasible and profitable. The potential *benefit* was clear, to receive the property tax exemption, which meant achieving the charity care requirement, but the real *cost* to the potential cost of reaching that goal had no such clarity, in large part because many of the same personnel and equipment were used to treat both rich and poor patients, with or without health insurance, and with varied health conditions. But here were more complexities.

Before Provena, little or nothing hinged on a hospital's reported provision of charity care, for the property tax exemption did not depend on that particular service provision. The State's Department of Finance was not incentivized to monitor a nonprofit hospital's service-reporting accuracy and integrity. But after the Court decision, a hospital's accounting choices did matter. It became necessary for the State to monitor each hospital's provision of charity care, or to accept its claimed provision, and to prevent gaming of its reports, by "cooking the books," and changing accounting practices, legally or not, to reach the charity- care threshold and retain the property tax exemption. No small task.

But the goal of Illinois regulators was different -- to confront nonprofit hospitals with stronger incentives to admit and treat more needy patients and treat them more "fully;" the *measure* of their "successful" performance was not some vague concept of treating the poor. It was the *dollar amount* of *charity care* reported to the state by the hospital. That number was not merely a tabulation of the hospital's intake, in a given year, of patients who were poor and uninsured; it depended on the hospital's valuations of the resources it claimed to have devoted to their care. That depended on whether care for a nonpaying patient was self-reported as *charity care* or as *bad debt*; and that distinction was often unclear.

The title of a recent article captured at least some of these measurement complexities: "Is It Bad Debt or Charity Care? The Right Way To Measure Uncompensated Care." [26] But is there a "right" way? And what does that mean? What ways are there for an organization that is rewarded for "better performance" to find ways to "beat the system"?

The size of a hospital's bill to a poor patient permits hospital choices and provides incentives. Much of the cost of operating any hospital, school, college, police department, museum, or court, for

[26] Jones, "Is It Bad Debt Or Charity Care?"

example, consists of overhead, "joint," costs that are not attributable to any specific consumer or illness. This goes far to explain why a tube of bacitracin antibiotic ointment—which, as one medical analyst put it, "…my mother put on the scrapes I got as a kid and that cost $5 at CVS" was priced at $108 by at least one hospital, and why "a three-hour emergency room evaluation for chest pain caused by indigestion" brought a [hospital's] claim of $21,000 for charity care." [27]

According to the Federal government's Center for Medicare and Medicaid Services, there remain vast differences among hospitals in their charges (prices) for treating a patient with a particular illness diagnosis. The differences are not random numbers; they reflect decisions by hospitals pursuing their own objectives, subject to external constraints and incentives. It was alleged that the Bayonne Medical Center, in New Jersey, for example, typically charged $99,689 for treating each case of chronic lung disease, 5.5 times as much as other hospitals and 17.5 times as much as Medicare paid in reimbursements. The hospital also charged an [sic] average of $120,040 to treat transient ischemia, a type of small stroke that has no lasting effect. That was 5.8 times the national average and 23.6 times what Medicare paid.[28] And a phone survey of charges for hip replacement surgery produced prices ranging from $11,100 to $125,798 for uninsured persons.[29]

When *valuing* expenditures on "charity care," a hospital has wide latitude. Many choices can have major effects on the amount of expenditures reported as charity care. It may be quite easy for a hospital to increase its *charges* for particular services, especially those most likely to be provided to poor patients, and then record the dollar value of charity care at the increased prices. This is analogous to valuing private manufacturers' donations of goods to charities by their "list" prices, even though widespread discounting means that it is the rare consumer that actually pays those prices. If charity care is valued at the list prices of its components, a hospital could "use the higher prices when calculating the amount of charity care it was providing …."[30] An instant increase in reported charity care would result, even though absolutely nothing of substance changed. Magic by gaming!

There is always another option: Lobby the legislature or otherwise work to rescind the law. For hospitals receiving or fearful of receiving multi-million-dollar property tax bills in a single year, and expecting it to continue for the future unless the hospital increased its "expenditures" on charity care – however ambiguous that term is -- changing the law had powerful appeal. Hospitals saw their social contribution as multi-dimensional, involving much more than providing free care to the uninsured poor, and so they widely favored the state's relaxing, if not repealing, the tight link between incurring more costs for charity care, and losing their property tax exemptions. The notion that there was nothing else that governments could reward was novel and, to the hospitals, seemingly arbitrary and costly. If only the law were changed, reversed to its prior, far broader, form, or at least tied to a wider range of public services than only charity care by a nonprofit hospital, however those terms were defined and measured. In fact, the law was reversed.

---

[27] Welch, "Diagnosis: Insufficient Outrage"

[28] Creswell, er al., "Bills at Hospital in New Jersey"

[29] Rosenthal, "The Price for a Hip Replacement"

[30] Creswell, et al, "Bills at Hospital in New Jersey", quoting Gerard Anderson, director of the Center for Hospital Finance and Management at Johns Hopkins University.

All or Nothing: Incentive Effects of "Pass-Fail" Performance Measures, Not Just for Hospitals

There was, and is, a feature of the Illinois property tax exemption policy for nonprofit hospitals that was not attacked either each hospital received *complete* exemption from *all* property taxation, or it received *zero* exemption, having to pay the full tax, as would a for-profit firm or residential property owner.

The incentive effect**s** of "pass-fail" reward systems are common but inherently inefficient. They only affect incentives when the taxpayer's performance is "close" to the threshold between having to pay the full tax and paying none. In the hospital property tax case, a modest increase or decrease in its provision of charity care could have a large financial effect if it either caused the total loss of *all* its property tax exemption, or it reinstated a previously withdrawn exemption.

The same strong incentive to surmount such a hurdle and reap a substantial reward in the process occur in other settings and industries. A student who is only a few points below passing a course or receiving an award, or a police officer who is one speeding ticket below the monthly quota, or a low income family that would lose its eligibility for an anti-poverty program if family income increased by a few dollars per month, would face a similar "pass-fail" incentive.

What about the hospitals, schools, police, and the poor who are already considerably above, or below, a cutoff threshold? There would be *no* financial incentive for them to provide any additional charity care, or for a student and parents to agonize over a student's standardized test score if only passing or failing mattered and the student scored considerably below or above it; there would then be little or no incentive for the stakeholders to try harder to improve their performance on standardized tests, or for police to work harder to increase arrests of marginal auto speeders, or for a low-income family to decline an opportunity to earn additional income lest that cut or even eliminate the family's eligibility for anti-poverty aid.

Pass-fail incentives characterize much of the public and nonprofit sectors, generating a variety of inefficient incentives. The essence of the inefficiencies is that it is rarely true that there is a unique level of performance that has great social significance but unless that threshold is reached, *better* performance is treated as unimportant, and once the threshold is reached, further improvement brings no additional reward, so setting rewards to incentivize *only* a "passing" grade is typically unproductive. A more efficient incentive structure would reward better performance more generally, and would expand the rewards to other government and nonprofit activities such as museums, philanthropies, courts, and more, for exceeding an essentially arbitrary performance quota; after all, the incremental social value of *better* performance does not suddenly become zero when a quota is reached or surpassed.

Why, then, would government and nonprofit sector providers so often be faced with strong rewards for meeting or exceeding some threshold of performance, but little or nothing *additional* for surpassing that level, nor for *improving* performance but remaining below the threshold?

Threshold effects are common. Apart from hospitals they are used often in education, where a point or two on a standardized test in New York State can have major effects on whether, for example, a high school student graduates with a distinguished "Regents" Degree. Since a near-miss might well affect college admission prospects, a high-stakes test can set in motion responses that weaken student applications, such as by having exams initially graded as close to meeting the Regents' award level being re-graded somewhat more liberally by another teacher.

Quantitative scores are broadly appealing because they facilitate rankings governmental or private nonprofit institutions. At the same time, they are troubling in their magnifications of what may be "small" and even inconsequential differences.

In the hospital case, of the choices available, option 4 prevailed: the law was changed. Rather than pressuring hospitals to increase their charity care, with all the incumbent complexities and distortions, the ways a hospital could qualify for property tax exemption were expanded. On May 29, 2012, the "Medicaid reform" bill passed both houses of the Illinois General Assembly and listed not only charity care but six other forms of "qualified services" that would justify a hospital's exemption from property taxation. While all addressed ways that a hospital was either providing health services to low income and underserved people, or otherwise was relieving the fiscal burdens on state or local governments, more activities would now count toward qualifying a hospital for property tax exemption services to the poor, but would no longer be required to be entirely free to the patient.[31]

For the hospitals, this was certainly the least costly, most desirable, solution. It required no change in their services, no additional charity-care services or other expenditures to regain their property tax-exempt status. The Illinois legislature improved hospital "performance" by the stroke of a pen -- expanding the definition of performance, its measurement and rewards.

The tying of tax-exemptions to provision of a specific service to a particular population group no longer holds, let alone being linked to charity care. What has not been resolved, however, is the wider issue of whether exemption from a property tax or any other tax *should* be contingent on provision of specific measurable services to specific beneficiary groups:

- Insofar as a particular tax-approach worked to stimulate provision of a public service, why not extend the principle to other organizations and services? Why not include service providers that have significant untaxed assets in the form of endowments, or that receive substantial benefits from tax-deductible private donations?
- For a *college*, why not increase its measured performance by making it easier for students to "pass" standardized tests, and encourage grade inflation that increases GPAs and the graduation rate, common measures of a school's performance, and increase college basketball players' graduation rate, simply by changing the way it is calculated. There is more about these forms of gaming when we turn to the paper (in this series) on higher education.
- For a police department's "performance," the reported amount of "crime" can be reduced by changing reporting practices. What, after all, is a law violation? In a memorable case a local police officer reported receiving a call claiming an alleged "theft" of a woman's purse, but after a discussion the caller agreed to its being reported as a "mysterious disappearance." Not a "crime," but even if it were, it would be difficult to value.

[31] Frost, Peter, "Legislation Defines Charity Care for Hospitals."

Succeeding Without Really Succeeding: Speeding College Basketball Players' Graduations

The easiest way to improve *performance*—as it is measured--is to game the measurement process, changing the way performance is measured. To say that one way is "easiest," though, is not to say it is easy, let alone costless in terms of the resources required or adverse side-effects; it rarely, if ever, is. But it is to say that efforts to "improve" performance of public service markets will confront obstacles as decisionmakers search for the most profitable path to the rewards, including changing the way performance is defined, valued, and measured.

Seldom has the difference between success in improving *real* performance, and as it is *measure*d, been clearer than in national efforts to improve the *academic performance* of college basketball players in the "big time," Division 1, universities. These "student-athletes" have had notoriously low college graduation rates, which have been interpreted as reflecting the emphasis on athletics, winning games, and generating revenue from commercial endorsements by producers of athletic equipment and clothing, rather than q players' academic accomplishments.

Multiple goals --in this case, to develop college students' academic as well as athletic skills, to win games and to generate revenue for the university -- are common in governmental and private nonprofit realms, and not by chance. By contrast with private profit-maximizing firms, these organizations combine public subsidies, on one hand, with legal restrictions and mandates that attract producers that pursue multiple goals, including provision of collective ("public") outputs such as cleaner air, aid to the homeless, and opportunities for individuals to supply time as volunteers. Getting all these incentives "right" for administrators, teachers, coaches and players is a pervasive challenge. How, then, might universities go about changing incentives to optimize the multiple and often changing goals, athletic and academic, collective and private?

One answer: change the ways performance in the varied dimensions is *measured*. This will become clearer when we examine, in a later paper , the route the National Collegiate Athletic Association (NCAA) took in response to criticism of its college basketball players' "low" graduation rate? It changed how the rate was *measured*.

Success in *winning games* was easy to observe and measure. But increasing basketball players' "graduation rate" turns out to be a form of *performance* that is much more subject to debate, gaming, and manipulation than we might think; there was no "right" way. Nor was there a right way to allocate rewards for overall success, since that is often a result of teamwork. In June of 2021 the U.S. Supreme Court ruled that college athletes have property rights in their own skills and images, and so may be paid for their market value, which encourages players to do what maximizes their private value, not that of the team or its players, coaches. and publicists.

For college basketball, a hard and costly way to increase players' graduation rates and their scholastic GPAs would be to cut their athletic practice time, increase their academic study time, increase tutoring as needed, and otherwise encourage student-athletes to pay less heed to winning games and more to academic studies. Universities could give coaches contracts offering larger bonuses for increasing players' graduation rates, even if that meant sacrificing victories on the basketball court. Coaches could be encouraged financially to change recruitment practices to attract basketball players who are more interested in academics and graduating. Schools could end the current practice of accepting academically sub-par students as "special admits" because of their athletic prowess despite

lack of academic skills and motivation.[32] These are all possible, but at a price: increased costs and decreased wins.

Incentives Matter, but Getting them Right—That is the Challenge

Colleges certainly could change their coaches' contracts and incentives, but not without consequences. Changing rewards from winning games to advancing players' academic accomplishments for graduation is possible but unrealistic at Division 1 Universities.

A study of *football* coaches' contracts at Division 1 universities compared their opportunities to earn bonuses for winning games and for speeding players' graduation rates. It made clear what those universities have seen as their football coaches' central mission. Thei contracts are striking: There were and are far greater bonus opportunities for coaches who win more games than for coaches whose players' make greater academic accomplishments. For football coaches at eleven major (Division 1A) conferences the maximum bonuses a coach could receive for good performance in winning additional games and for players' graduation success -- higher team GPA and six-year graduation rate—left little doubt about what matters most to school administrations.

In all of the major conferences across the country, coaches had substantially greater bonus opportunities for winning games -- ranging from a low of three times the bonuses available for academic accomplishments, to highs of 37 or even more times the maximum academic bonus. Moreover, the differential potential monetary bonuses ranged from a low of $22,000 at Kent State University, to highs of $275,000 at the University of Louisville, $681,000 at Auburn University, and $940,000 at the University of Virginia; the largest relative bonuses for winning games were at schools ranked highly nationally in their football achievements.[33]

Grade inflation in classwork is another way to speed the graduation process, and not only for athletes. A former academic colleague of mine told me that when he began college teaching. in the 1970s, there were five grades in undergraduate courses: A, B, C, D, and F. When he retired, forty years later, there were still five grades—but they were: A, A-, B+, B, and B-. Well, not officially of course-- the C, D, and F grades remain-- but their relative frequency has plummeted. And there is little doubt in the academic community that grading standards have been lowered.

Redefining a university's graduation rate, at least as applied to athletes, is what the National Collegiate Athletic Association (NCAA) did in 2007, in response to pressure to do something about the abysmal graduation rate of Division 1 basketball players. No longer was graduation "success"—performance-- measured by the percentage of team members who graduated within 6 years; it was redefined to *exclude* players who transferred from another school, but to *include* incoming transferees who graduate from any school within six years. Termed the Graduation Success Rate (GSR), this change in measurement dramatically "improved" graduation performance.

The lesson is powerful: whatever the industry or context, if the rewards to any stakeholder hinge on performance as it is measured, stakeholders will have the incentive to develop and implement new measures to generate maximum rewards. But there will be adverse side-effects.

---

[32] Weisbrod, Ballou, and Asch, *Mission and Money*, 232-234.
[33] Weisbrod, Ballou, and Asch, *Mission and Money*, chapter 14.

Not surprisingly, the U.S. Department of Education's longstanding collegiate definition of the 6-year graduation rate, and the NCAA's newer GSR, differed sharply. For student-athletes who enrolled between 1997-98 and 2000-2001 the two graduation rates for men's basketball players at the University of Florida, for example, the GSR was a perfect 100 percent—every team member graduated within 6 years; but when the larger number of players counted in the Federal government's calculation was included, only 67 percent graduated. In another sport, baseball, the NCAA's measure showed a 6-year graduation rate of 71 percent, compared with the Federal government measure of only 26 percent. [34] The differences between the governmental and NCAA measures of success were *no*t the results of either differential accuracy of the measures or of honesty of reporting, although incentives to exaggerate performance – to cheat in order to increase compensation - - is an ever-present danger, throughout the economy.

Pay-for Performance in K-12 Schools: the Art of Cheating

In 2009 the Superintendent of Schools in Atlanta, Georgia, Beverly L. Hall, was named "Superintendent of the Year" by the American Association of School Administrators. She received national attention for her leadership performance, and U.S. Secretary of Education Arne Duncan hosted her at the White House. Ten years earlier Dr. Hall had arrived in Atlanta, taking what would be her last step up the education administration ladder -- first through the New York City schools, as a teacher, principal, and deputy superintendent, and then as superintendent in Newark, New Jersey, before Atlanta. She had "built a reputation as a person who got results, understood the needs of poor children and had a strong relationship with the business community." Under her leadership, performance of the Atlanta Public Schools improved sharply, as measured by the percentage of students who passed the standardized tests. They soared.[35]

The chief operational measure of Superintendent Hall's successful school leadership performance was clear—more students "passed " the standardized exams. The financial rewards for her success were equally clear: the link between student performance, as measured, and *her compensation,* brought her over $500,000 in bonus P4P rewards, in addition to the national award. There were also rewards for her supporters, the teachers and school principals whose students' test scores were escalating; there were more job tenure appointments and more performance bonuses.[36]

Fame and its rewards were short-lived. Within two years, in 2011, her fame disappeared. Superintendent Hall was out of a job, a number of teachers had been discharged, and a Grand Jury had indicted her and 34 teachers, principals, and administrators of the Atlanta Public Schools, on charges of "racketeering, theft, influencing witnesses, conspiracy and making false statements." Prosecutors recommended a $7.5 million bond for Dr. Hall, and convictions on all counts could have led to 45 years in prison.[37]

Hall "retired" amidst claims of systematic cheating. Teachers and principals had allegedly erased students' wrong exam answers, substituting correct answers. Georgia State investigators' 800-page

---

[34] Weisbrod, Ballou, and Asch, *Mission and Money*, 238-239.
[35] Strauss, "How and why convicted Atlanta teachers cheated"
[36] Winerip, "Ex-Schools Chief in Atlanta"
[37] Winerip, "Ex-Schools Chief in Atlanta"

report implicated 178 Atlanta teachers and principals, as well as Hall, including 82 who confessed.[38] Under Hall's leadership, one middle school principal in his very first year, in 2005, oversaw the percentage of his eighth graders who scored "proficient" in math more than triple, leaping from 24 percent to 86 percent. Investigators found cheating in 44 of the 56 schools examined, and concluded that Dr. Hall "should have recognized the cheating because the gains on the tests were so swift and dramatic."[39]

The cheating, if confirmed, was certainly improper and in many cases illegal, but was it perilous, systematic, counterproductive, or even worse? Yes, it was.

Or was it just the result of a few "bad apples" in a school system with thousands of teachers and administrators? No, it was not; more was involved than some teachers, principals, and the Superintendent being rewarded for bogus achievements.

Because the incentives were perverse, because they rewarded falsified answers and led to inflated test scores that exaggerated improvements in students' true learning, they led to a price being imposed on other stakeholders. The Parks Middle School, which appeared to perform so spectacularly in terms of standardized tests, was deemed to no longer need improvement, and so it lost $750,000 in state and federal aid that could have been used for the many students who continued to need extra academic help. A teacher at Parks Middle told investigators that she had students who scored "proficient" on the falsified state tests but who actually could read only at the first-grade level.[40] Those students were collateral damage imposed by the teachers and school administrators who had elevated their own measured performance and its rewards, but in the process lost governmental financial aid that had previously been available to help low-performing students. So, the rising class-average performance data that was generated by the cheating came at the expense of the low performers, who lost their supplemental assistance.

There were more side-effects. An Atlanta third-grade teacher reported that cheating had been overseen for years by the school principal, "who wore gloves so as not to leave fingerprints on the answer sheets." [41]

Evidence of adverse effects of strong rewards for "good performance," as it is measured, is by no means limited to schools, but can distort incentives and choices wherever managers and employees in any industry can take advantage of their "inside" information. So whatever the industry, and whether the competition involves for-profit companies, government agencies, or private *nonprofit* firms and employees, there can be, and often are, ways to purposefully overstated productivity, to increase their performance-based compensation, albeit illegally gaming, by employees falsifying their own performance, has been engaged-in by police law-enforcement personnel—who recognized that they were better informed than were their supervisors.

---

[38] Winerip, "Ex-Schools Chief in Atlanta"
[39] Bowers, et al., "Office of the Governor: Special Investigators."
[40] Winerip, "Ex-Schools Chief in Atlanta"
[41] Winerip, "Ex-Schools Chief in Atlanta"

Police Performance and Incentives: Goal Conflicts and Gaming

In 2010, the Chicago Police Department took steps to improve its anti-crime performance, by incentivizing its personnel in two ways. Specifics of the reward and penalty structure, and the specific bonuses paid for "better performance," were not made public, but two program goals appear to have affected police rewards through bonuses, promotions, and other awards for advancing Department goals – specifically by (1) cutting police use of racial, cultural, ethnic, and religious "profiling" as means for identifying likely law-breakers and reducing crime, and (2) reducing police incentives to detain, question, and perhaps arrest, car drivers, pedestrians or others individuals identified as behaving "suspiciously," an ill-defined term allowing police discretion if not arbitrary actions.

Police were, in effect, given two goals, which conflicted: (a) *apprehend* more law violators, but in the process, (b) *avoid using profiling to identify more-likely violators* by virtue of their visible characteristics such as skin color, ethnicity, religion, and clothing.

Some police officers devised a creative "solution" to deal with the conflicting goals and their incentives -- how to maximize police stops, frisks, and arrests, while minimizing use of profiling tactics that are inherently discriminatory. The police had been told to avoid profiling in their selection of people to "stop and frisk," but they were also told to identify law violators, and the easiest way to do that may well have been to profile.

Monitoring the anti-profiling policy thrust police into a new double-edged role of collecting detailed information about suspicious behavior in the area, but not engaging in racial or cultural profiling to verify it. Henceforth, whenever a police officer stopped someone for questioning, a specific reporting form was to be filled out by the officer. The form asked for the racial and cultural characteristics of the individual detained, whether any evidence of illegal behavior was found after the stop, and whether the person was formally charged. An officer who consistently detained "excessive" proportions of, say, Blacks or Hispanics, could be penalized, yet the magnitudes of deterrent effects is not clear. The challenge in such situations where police have two or more roles that conflict – in this case, advancing the *public* interest in reducing crime, and the *private* self-interests of individual police men and women -- are how supervisors can distinguish between *random* errors of police reporting of stops, and systematic biases by police in choosing who to detain .

The challenge of incentivizing to achieve conflicting goals is compounded by information inequalities, well-illustrated here, in which the people with the most information—in this case the police patrols -- do not have the incentive to reveal to their supervisors' details of their performance if that would be penalized, which was almost surely the case. The patrols have an informational advantage that conflicts with their self-interest in gaming the rewards for stops, detentions, and even arrests, while avoiding penalties for added profiling. The incentives made sense: encourage detentions and even arrests, but discourage profiling. Or so it seemed.

Good Intentions and Good Results: Multiple Goals in Policing

Whenever there are multiple goals, problems can be expected to emerge for establishing and enforcing efficient rules, rewards, and penalties. Moreover, and this is not only common, but typical, whenever rewards and penalties in some forms are less costly to measure accurately than are other forms, the more distortionary the reward system will be. This insight was developed in research by economists Bengt Holmstrom and Paul Milgrom in 2016, in the broad context of incentive-based

employment contracts – employers and their employees, or higher and lower-level personnel. That model was at work in the Chicago police department.

The top echelon of police officials had apparently not foreseen the potential for lower-level police to game the reward system, which provided incentives to cheat on their reports of accomplishments, and, in turn, to bring acclaim, bonuses, salary increases, and promotions. Apprehending more law violators, for example, could be advanced by "profiling" probable law-breakers – that is, by identifying people as likely to violate laws, based on his or her ethnicity, religion, skin color, or clothing. But profiling, especially racial, has been widely condemned on discrimination grounds, for favoring or disfavoring people with particular physical characteristics.

Police have developed law-enforcement mechanisms showing progress toward their goals. Evidence of progress was submitted in reports to supervisors, showing, for example, that more car drivers who police stopped and detained had been driving "suspiciously," indicating that stricter police enforcement was working. But at the same time, there were *fewer* such police reports of their "stops-and-frisks" of minorities, particularly Blacks, which were deemed to be evidence of improved police relationships with *minorities*.

Racial profiling was not actually being halted; neither was it being materially diminished. Police were simply *falsely* reporting having made *fewer* stops of Blacks and *more* of Whites than actually occurred, even though the truth was that police were gaming their own performance reports, to support claims of favorable, though false, accomplishments. So, two police performance goals were seemingly being advanced: (1) police use of racial profiling was allegedly decreased, as the proportion of persons who were "stopped and frisked," and then reported as being Blacks, was intentionally *under*-reported, to show decreased police use of racial profiling against Blacks, and (2) the proportion of detainees reported as being *White* was being *overstated. Both forms of misreporting were not accidental; both were intended to narrow the gap in police reports of their handling the various racial groups equally.* But how to do it?

By cheating, police submitted overstated reports of their own achievements to their supervisors. Police on patrol worked out a clear and effective reporting strategy: they took advantage of the high cost to their supervisors of verifying the accuracy of the reports being turned in. Those reports appeared to show that police use of racial profiling against Blacks was decreasing; but that was fiction, the result of intentional under-reporting of numbers of "stops-and-frisks" of minorities. And factual evidence was being falsified to mislead police supervisors, by systematically mis-stating key evidence from their car-stops. Reports to supervisors did produce evidence of law violations, but the racial characteristics of the violators were systematically mis-stated. When *White* detainees were found to be *innocent of any legal wrongdoing, they were* sometimes being reported as being *nonwhite*. When *nonwhite* detainees were determined to be in violation of a law they were sometimes reported as *White*; the systematic falsifications thus *over*stated the proportion of Whites who were reported as law violators, and *under*stated the proportion of "nonwhites" reported as violators. So the reports disclosed greater equality of probabilities of detention of innocent whites and nonwhites than was actually the case. [42] Police self-reporting disclosed, but incorrectly, their adherence to official police regulations.

---

[42] See Mahr and McCoppin, "Racial Mislabeling" and Luh, "Misreported"

"Gaming" in Policing and Public School Teaching: More Alike Than Not

The systematic misreporting by police of their performance is reminiscent of the misreporting by k-12 students of their grades on standardized tests, in the Atlanta public schools, under the No Child Left Behind (NCLB). Act. Teachers and administrators were erasing students' wrong answers on the tests, and replacing them with the correct answers, which the school system rewarded.

While both the Atlanta schools and the Chicago police deliberately cheated on reports used to reward *performance*, the police misreporting highlighted another characteristic of pay-for-performance, the likely informational advantage of some stakeholders over others, and the incentives open to the better-informed teachers, school administrators, and lower-level police to benefit from using their informational advantages to game their higher-level but less-informed supervisors.

Taking advantage of such informational superiority to reap a reward for "good" performance can, but need not, involve illegal behavior; it does, however, permit teachers, hospital administrators, and police, for example, to utilize regulatory "loopholes" to garner private rewards for intentionally overstated performance. The better-informed stakeholders acted in their private interests at the expense of the less informed.

Measuring and rewarding performance by police highlights a distinction from incentivizing k-12 teachers: In policing there is no equivalent of the standardized tests in schools to gauge performance by individual police. The result was stronger police incentives to make decisions and reports based on their own *judgments* on whether to make "arrests" and how to report a detainee's race. The school test-cheating scandal, by contrast, involved cooperation between police supervisors and their subordinates, to correct students' wrong answers.

In February of 2012, the Washington, D.C. Police Department confronted the consequences of its multiple goals and the challenges of measuring achievements. Homicide detective Milton Norris won the Homicide Detective of the Year award for his success as measured by his crime "closure rate"—the percentage of all his cases that were closed during the year. Measurement of success was critical.

The officials who gave Norris the award announced that he had "closed" all of his recent, year 2010, cases. But a review of his homicide cases by *The Washington Post* found that he had a far less stellar performance record; he had only two cases that year, and both remained open at year's end. Further investigation disclosed that Norris had received credit for closing several cases from previous years. In reality, his performance had been mediocre for the preceding six years between 2004 and 2010; he had investigated 41 homicides and closed 24, a 58 percent closure rate, according to police records. About 10 of his cases ended in convictions, court records show, and three were closed with no arrest.[43]

Understanding the Pay-for-Performance Problem: It's All About Measurements

Adam Smith had recognized the efficiency appeal of pay-for-performance nearly 250 years ago, in 1776; rewarding producers for satisfying consumer demands aligned their self-interests. In today's mixed economies, the principle is unchanged; teachers' and students' interests need to be aligned, as do

[43] Thompson, "Mystery over cases"

hospitals' and patients' interests, police supervisors' and their subordinates' interests, and so on. More generally, social efficiency requires that the interests of organizations, their employees, and consumers need to advance together.

When performance is multi-dimensional, all forms of performance must be measured, valued, and rewarded. Otherwise, bias can result if any form is consistently over-or under- rewarded relative to others. Given the measurement problems, however, a comprehensive reward structure is seldom feasible. What happens when some forms of productivity are observable and are rewarded, while others are not? *Tha*t is what makes *strong* rewards for *measurable* performance perilous.

If a school's goals include educating students in the 3-Rs but also in social studies, music, and art, it would want to measure a teacher's contribution to each, and then turn to the valuations and methods for converting the values into rewards to the responsible stakeholders. The obstacles to implementing these rewards are especially likely for goods and services supplied by government and nonprofit organizations, because those organization forms and their public and private constraints after determining the values of incremental successes in each subject, rewarded each form and increment of performance. Without objective measures of performance in each dimension—no small task -- especially in the public and nonprofit sectors—and mechanisms for valuing additional performance productivity of a teacher, school, hospital, museum, police, or any other agency's contribution, independently and jointly, measuring the value of advancing the system's goals will not merely be incomplete; it will be biased downward, with too little incentive for a teacher, hospital administrator, police department, museum, philanthropy, or other stakeholders to devote effort to the *un*rewarded or poorly rewarded forms of performance, whether the lack of rewards and incentives resulted from measurement problems or absence of rewards.

Governmental and Nonprofit Organizations are Not Just Like Private Firms

Goals differ among governmental, nonprofit, and for-profit organizations even though they are alike in some ways. Mixed ownership forms coexist in many industries including higher education, health care (hospitals and nursing homes), and policing. Among hospitals, public and nonprofit providers have been found to be similar, but both differ from for-profits, providing, for example, a markedly wider range of services, many of them presumably unprofitable "mission" goods for the publics and nonprofits that for-profits avoid.[44] Because of the particular activities that governmental and nonprofit organizations engage in, rewards for their good performance, and the associated incentives, differ.

Even churches—which are also nonprofits--are not immune from the pressure to reward their personnel for forms of performance that are easily observed. Over the 43 years, 1961-2003, financial compensation of more than 2,200 Methodist ministers at all 727 United Methodist Congregations in Oklahoma were paid for their performance, as measured by Congregation total membership; more members meant higher pay.[45] There was no direct evidence, though, of causation; was the quality of a minister's "performance" actually measured by membership, or were other outcomes relevant ?

[44] Weisbrod, "How Mixed Industries Exist"
[45] Hartzell, et al., "Is a Higher Calling Enough?"

When any organization -- a hospital, school, police department, church, or anything else -- faces opportunities likely to bring larger rewards, there are perils; will those rewards reflect truly improved performance or only *pseudo success* that overlooks unmeasured or biased effects?

Schools are especially susceptible to gaming, perhaps because of the complexity of gauging their performance for a diverse student body. In El Paso, Texas, the school superintendent was convicted and imprisoned for "removing low-performing children from classes to improve the district's *average* test scores." In Ohio, state officials have investigated whether several urban districts intentionally listed low-performing students as having "withdrawn," and so they were excluded from testing, although they actually were attending school.[46] In both cases, when low achievers were dropped from the test pool the average scores on standardized tests rose; *measured* school performance went up, not because the quality of education had improved but because weak students no longer reduced the class test average.

The Pay-for-Performance problem is far more fundamental, as individuals and systems adjust to incentives not to the *true* goals but to the actual ways that teachers and administrators are rewarded; that depends on how their performance is *measured*. In France, where, since 1808, a student could not graduate from high school and go on to college—the *reward* -- without passing a national test of his or her educational proficiency (*performance)*the "bacs" (*bachelliers*) test.ng," have been weakened enormously, thereby making it easier to qualify for the heavily subsidized higher education. As recently as 1945, high school graduation was "the province of an intellectual elite, held by only 3 percent of all teenagers," but by 2012 bacs were passed by 80- 90 percent.[47] A similar phenomenon occurred in America, as less than 25% of the population aged 25+ had completed high school in 1940, but that increased to over 90% by 2020.[48]

Since earning a "bacs" is necessary for college admission in France, and since a college education has increasingly come to be seen as the doorway to economic and social success, the incentive for schools to relax standards for passing has been powerful. Despite having passed the bacs, however, fewer than half the first-year students at French public universities continue to the second year. Erosion of standards for the bacs is allegedly a major culprit: According to one political leader, "Rather than raise 80 percent of students to the level of the test, "we put the baccalaureat [sic] at the level of 80 percent of students.'"[49] Grade inflation followed.

Erosion of academic standards is not harmless. Neither are the distortionary incentives to game the reward system—for increasing student learning, reducing crime, increasing hospital care for the poor, or anything else. When a school principal, hospital administrator, or police officer is rewarded for performance that is systematically *mismeasured,* the incentives are inefficient. Even if unwittingly and unintentionally, they reward and encourage actions that do not advance the public interest.

The systematic problem, though, is mismeasurement of social performance: private incentives reflect compensation only for the measured, observable, forms of performance, not for unobserved

---

[46] Winerip, "Ex-Schools Chief in Atlanta"
[47] Sayare, Scott, "Rite of Passage for French Students"
[48] US Census Bureau, *CPS Historic time Series.*
[49] Sayare, "Rite of Passage," quoting Jacques Juilliard, columnist the newsmagazine Marianne.

forms. As a result, private rewards and social performance are misaligned. Unmeasured elements of social performance are essentially ignored.

This process is visible throughout the public and nonprofit sectors. School administrators and teachers focus curricula on reading and math, as measured by standardized tests, and cut back on time devoted to physical fitness and the arts, for example, which are little rewarded. Hospitals, nursing homes, police—all focus their resources on the performance elements that bring them rewards. Charities, similarly, pay attention to their easily measured data such as their financial "efficiency" ratio—funds raised per dollar of fundraising expense—even though that ratio conveys no information about effectiveness of the programs on which the funds are spent.

With performance" being increasingly rewarded through governments and nonprofits, the incentive to find shortcut, low cost, ways to demonstrate success—to measure it and qualify for greater rewards, is accelerating. Changing wrong answers on student tests is faster and cheaper than improving teaching or changing students' study habits. Changing a hospital's accounting practices to show greater "expenditures" on "charity care" is cheaper than attracting and caring for more poor and sick people. For a charity, it is easier to alter its accounting practices to show greater fundraising "efficiency," than to improve program effectiveness.

Claims of greater accomplishments by nonprofits and governmental agencies abound; evidence does not. Unsupported claims that a program "creates relationships among children, students and parents that are rarely found anywhere…" often suffice to generate program support. As the recent CEO of one large nonprofit, National Public Radio, Ken Stern, stated: "For most charities, … the most important measure of success [is] one that typically confirms the importance of the work and reassures stakeholders. Empirical and research studies are to be avoided as expensive, distracting, and potentially dangerous." [50]

The limited role of performance measurement in justifying social programs is illustrated in the U.S. by a much-publicized youth anti-drug program, D.A.R.E. (Drug Abuse Resistance Education). This program, initiated in 1983 by the Los Angeles Police Department in partnership with the nonprofit Rotary Club of Los Angeles, focused on students as early as fifth grade, had a curriculum emphasizing the dangers of "drugs," both illegal and legal—including alcohol and tobacco. Its program legitimacy was enhanced by having all the D.A.R.E. teachers be police officers who had received "at least eighty hours of specialized training in child development, classroom management, elementary school teaching, and communication skills." [51]

Evaluating the *performance* of any program requires determining its costs as well as benefits. The resource costs of the D.A.R.E. program, including the thousands of police officers involved, has been estimated by D.A.R.E. at about $200 million per year in Los Angeles, and, nationally, where some 7-8 thousand full-time equivalent police officers have been involved, at between $1 and $1.3 billion annually. Moreover, as Stern summarized the program's effects, even apart from the costs, he

---

[50] Stern, *With Charity for All*, 37.
[51] Stern, *With Charity for All*, 31-32.

concluded "The weight of scientific evidence was overwhelming: D.A.R.E. doesn't work…." [52] Yet it continues.

Why Continue Governmental and Nonprofit Programs that "Don't Work" ?

Is it irrationality, or is there more? Much has to do with performance measurement and the rewards or punishments that accompany it. n a classical private enterprise market, weak performers lose customers, profitability drops, and firms exit. The principal measure of performance is profit; the penalty for poor performance is organization merger or bankruptcy.

In public goods markets, where governments and nonprofits dominate, unprofitability is not a clear indicator of poor performance. Unprofitability is expected. The unprofitability of U.S. Postal Service Saturday mail delivery is acknowledged but not necessarily treated as poor performance. Lack of success by the D.A.R.E anti-drug program has not brought it to an end –-apparently not signaling poor performance. Neither does the lack of commercial revenue mean the demise of the U.S. Department of Defense, nor does its unprofitability signal that military defense is not "performing well"—not valued at more than its cost.

One reason that a "failed" public or nonprofit activity might persist is inefficiency, perhaps resulting from corruption or poor transmission of information. Another is more complex: No matter how convincing evidence may be that a specific public or nonprofit program "does not work," it does not answer the question of whether some change in the program might be more successful. Rather than simply abandon an anti-drug addiction program, for example, the program might be modified or merged with another public service provider,

"Learning from doing," and modifying programs accordingly, can certainly be efficient, anywhere in an economy. So, too, can external events cause a well-performing program to become unproductive but not make it efficient for the organization to die. When highly effective polio vaccines were developed in the 1950s the rationale for the nonprofit National Foundation for Infantile Paralysis ("March of Dimes") soon disappeared -- mission accomplished. To the surprise of many, however, the Foundation did not die; rather, its target morphed from fighting polio to fighting other childhood diseases and birth defects. One rationale for the transformation rather than its simply closing its doors, was that the Foundation had become a quite successful fundraising vehicle for one serious disease of youth—an asset not easily developed and, arguably, worth preserving.

If the National Foundation for Infantile Paralysis had been a private firm rather than a nonprofit, its component parts that had value – in this case its fundraising infrastructure -- would have been recognized in the marketplace, as was the analogous case of Hostess Baking Company's 2012 bankruptcy and subsequent sale of rights to its "Twinkies" cake product. In effect, the Polio Foundation "sold" its well-performing fundraising arm to the new foundation addressing other diseases hitting youth.

[52] Stern, *With Charity for All*, 38.

Public and Nonprofit Organizations Differ in the Measurability of their Performance: the Road Ahead

The variety of measurement dimensions, the associated rewards, the nature of the perils of motivating better performance with greater rewards, and why the perils are especially problematic in the realms of government and private nonprofit organizations, are explored throughout the papers ahead: in k-12 schooling , health and medical care , higher education , policing, jails , museum, philanthropy), measuring performance quality through such mechanisms as ratings, rankings, and accreditations, and even Federal Courts and Judges .

All of these are parts of modern industrialized economies that are dominated by public and nonprofit providers rather than by private firms. Despite their differences, as well as commonalities, they bring measurement and valuation issues of wide-ranging scope; their social goals are multiple and generally difficult to quantify, their goals are multi-dimensional and difficult if not impossible to value accurately. Rarely can they be valued and summed to make them analogous to a private firm's profits.

These measurement and valuation problems stand in the way of public and private nonprofit organizations' efforts to establish reward systems that actually encourage "better performance." They are at the root of what permits any service provider, whether governmental, nonprofit, or for-profit, to "succeed" without really succeeding— why school teachers and administrators might correct students' wrong answers on tests to generate better school "performance," why students might intentionally do poorly on a test so they can show more "progress" on a subsequent test, why hospitals may discharge patients and readmit them a few weeks later, to receive additional payments from Medicare or private insurers, and why charities have incentives to shift the allocation of their *total* expenses between fundraising and program activities, so as to reduce the amounts reported to the IRS as "fundraising expenses"[53] and thereby show less fundraising expense and greater fundraising "efficiency."

These are many a few forms of strategic gaming of reporting systems to take advantage of regulators' informational handicaps, the high costs of their distinguishing better from weaker "performance," and the ambiguities that bring rewards even for poor performance.

The "Tyranny of Most-Popular Lists" as Measures of Performance

Opportunities for individuals and organizations to reap rewards for ostensibly good performance take untold forms. Consumers of countless products such as books, music, and movies, but also hospitals, nursing homes, and colleges, not to mention restaurants and cars, are rated and ranked to provide information-hungry buyers with evidence about what other people think are the "best." The reward for a "high" ranking may be direct—such as the $10,000 bonus feature in the 2007 contract of Michael M. Crow, then-President of the Arizona State University, if the University moved into the top 120 universities in the *U.S. News &World Report* ranking.[54]

Performance measures can be gamed, and they frequently are. Sometimes the gaming involves intentional falsifications . It is less costly to make it appear that a particular product or service is of high quality than it is to actually make it so . Showing that such manipulations of rankings can succeed, sociologists Matt Salganik and Duncan Watts ran an experiment on the effects of intentional *mis-*

[53] Taylor, Harold, and Berger 2013.
[54] Weisbrod, Ballou, and Asch, *Mission and Money*, 264.

rankings—in this case, of popular music--on consumers. In the experiment, one randomly selected group of subjects was shown a true ranking of the most downloaded popular songs, and another random group was shown a ranking in the reverse order, the least popular song being at the top. The subjects in both groups then downloaded songs of their own choices. Both groups did the most downloading of songs at the top of their lists; the subjects who had been shown the reverse list also chose to download the songs at the top of their lists, which were actually the lowest-ranked, least-downloaded, songs.[55]

The basic problems of understanding when incentives go awry also apply in private enterprise markets – measuring good performance and rewarding it appropriately -- and have been the subject of extensive study.[56] The intensity of the problems, though, are systematically lower in the private market, and for a somewhat surprising reason.

It is often claimed that governmental and nonprofit organizations are inherently inefficient compared with private firms because their managers cannot legally share in their organizations' profit—those organizations are subject to the "non-distribution constraint."[57] They are, it seems, less cost-consciousness and have less drive to develop profitable new markets, because successful efforts to achieve these efficiencies would be little rewarded. [58]

Whatever the validity of this perspective, it is important to recognize that the activities that are carried out by private firms and by government or nonprofit agencies, are systematically different from those of for-profit firms, for a good reason.

When consumers are well informed and their private benefits and costs coincide with the collective, societal, benefits and costs, the private market excels. When these conditions do not hold, it does not. In addition, those are the situations—reflecting private market "failures"—that give rise to government and nonprofit sector activity, where incentives to pursue private rewards vigorously are intentionally weaker.

Incentives and Informational Asymmetries

Getting incentives right is easy *if* the decisions made by providers—hospitals, schools, police, museums, charities, courts, or anything else—are rewarded for truly advancing social goals, and only for that. But these are big "ifs." The obstacle to establishing efficient incentives is overcoming informational *asymmetries* between buyers and sellers – as I noted earlier. These were focal points of the analyses by American economist George Akerlof, and British social worker Richard Titmuss in their independent 1970 publications. Both recognized that the parties to an exchange are often *un*equally informed about the quality and effectiveness of his or her "performance" – especially in public goods markets. Generally, though not always, it is the service supplier who has more knowledge than the consumer, client, patient, or employer, about the effectiveness and quality of whatever is being provided.

---

[55] Thompson, "Mystery over cases"

[56] For example, see Lazear, "Performance Pay and Productivity," also Holmstrom and Milgrom, "Multitask Principal Agent Analyses" discussion of job design.

[57] Hansmann, "The Role of Nonprofit Enterprise"

[58] Alchian and Demsetz, "Production, information costs"

When a repairman comes to a home to fix a malfunctioning cable TV, for example, he knows better than his employer whether the repair required his devoting as much time and effort as he actually took. But while the repairman knows, if he is paid by the day for making a specified number of house calls, he does not have the incentive to cut back on "wasted" time in order to increase the number of those calls. The employer could, instead, pay on a piecework basis, for each house call, just as an apple-picker might be paid according to the number of baskets filled. But *that would establish another disincentive problem—encouraging low quality work to save time and increase the number of house calls per day*.

The employer, whether governmental, nonprofit, or for-profit, cares about both quantity and quality of employee performance, but also recognizes that "quality" is often difficult to observe and so to reward. The dilemma is that the stronger the reward is for the more easily observed dimensions of performance, the greater is the worker's incentive is to disregard the unrewarded and the poorly rewarded, dimensions of performance; *small* rewards for the observables make sense.

The Problems with Pay-for-Performance: Measurements and Valuations

Tighter coupling of rewards with measured performance is perilous when despite the fundamental importance of incentives for better performance. In the realm of public demand for "accountability," evidence justifying a program's expenditures on the ground that it passes ~~a~~ benefit-cost test misdirect incentives if benefits and costs are systematically mismeasured or mis-valued. Pay-for-Performance rewards for public services may have as their rationale the stimulating of "better" performance; but frequently they achieve the opposite because the strong rewards apply only to a subset of performance measures, those that are most easily measured and valued. *Weak* rewards, by contrast, for poorly measured forms of performance can be efficient if the measures are costly and inaccurate, as appears to be the case of federal judges in the U.S. But why would performance be consistently poorly measured and biased, are important questions to be examined later papers.

To see the forces at work, consider efforts by law enforcement authorities to cut recidivism (re-arrest) rates of convicted criminals by using electronic monitoring rather than more-costly imprisonment. The key measurement issue is whether re-arrest rates after discharge differ for law violators who are randomly released from electronic monitoring compared with release after prison incarceration.

Researchers in Argentina studied released prisoners who had been randomly assigned among judges for sentencing—judges who differed in their ideological preferences for using each of the two forms of punishments, electronic monitoring (EM) or imprisonment. The researchers found a 11-16 percent lower recidivism rate for individuals subjected to electronic monitoring (EM) rather than imprisonment; the more effective option was also less costly.[59]

As the judges made their choices, what were the rewards and penalties they would consider? There appeared to be none. There were no differential rewards or penalties for Federal judges did not and still do not receive bonuses if the criminals they sentence had less post-prison recidivism. They have no financial incentive to use or not use Electronic Monitoring (EM). But with evidence of EM's greater

[59] Di Tella, "Criminal Recidivism after Prison"

effectiveness after imprisonment, as well as its *cost* advantage over incarceration, would it not be efficient to incentivize judges to make greater use of EM in their sentencing decisions?

Perhaps. There would be more use of the EM option and less use of prisons. However, since there was little if any evidence to support the conclusion that the lower recidivism rate was *caused* by the greater use of electronic monitoring, that conclusion would be premature. The randomized judges in the study made more use of EM even when they believed that incarceration would be more effective. Stronger incentives for judges to use EM would increase its usage and crowd-out incarceration, but implementation of that approach might well increase rather than decrease recidivism compared to judges' decisions previously based on their own assessments of the relative effectiveness and costs of each alternative. Do stronger incentives could be counterproductive, with Pay-for-Performance failing.

The message is sobering: No matter how well or how poorly performance is gauged, there are no consequences until the measures are linked to rewards. Stronger incentives are deceptively attractive. The problem is not simply mismeasurements, but our confidence in the connections between the performance measures and the rewards.

It's All About Incentives and Their Efficiency or Inefficiency

The thread connecting the diverse activities of a P4P reward structure is, in a word, *incentives.* The ways incentives operate, though, are both varied and often invisible to social planners. Strong or weak, stable or changing, all are products of an era of increased demands for connecting rewards with measured performance. And all are confronting the challenges – and the perils – of encouraging gaming of the measures. But there are many ways to deal with those challenges and opportunities.

The *perils* of pay-for-performance result largely from *inefficient* incentives. The riddle: why encourage antisocial behavior? The greater the rewards for improved performance, the more the distorted incentives encourage advancement of individuals' self-interest even when that comes at the expense of undermining social goals.

Bad incentives were at the root of the school scandal in Atlanta. Bad incentives were at the root of the Chicago police anti-profiling scandal. Bad incentives were at the root of efforts to increase Illinois nonprofit hospitals' provision of more medical care for the poor. In all these cases, and in others to be examined in later papers and in other governmental and nonprofit industries, a crucial problem is that efforts to improve performance by rewarding it more handsomely, often encourage unintended reallocations of time and effort. So, *measuring* performance often interferes with *rewarding* it efficiently, as stronger incentives encourage stakeholders to divert rewards to their private interests.

Within the U.S. Federal Government, in 1989, three members of its House of Representatives asked the Government Accounting Office "… to examine the government's pay for performance system … for supervisors and managers …." Most group participants were unhappy [but] they "had few suggestions for improving the system." The agencies studied were the Federal Aviation Administration, the Inrernal Revenue Service, the Office of Personnel Management, and the Bureau of Land Management. Participants generally believed that ratings were inequitable, and most "did not believe that performance was one of the major factors that determined who received performance awards." And these were just a few of the problems identified. And 33 percent of the employees and 23 percent of the supervisors believed that ratings were "not based on performance." (Fuhrmans)

The forthcoming papers in this series will examine a wide variety of industries, ownership forms, and the similarities and differences of their performance measurement problems, particularly the challenges of using large, high-powered P4P incentives. In the process I will also examine the case for using weak, low-powered, incentives, and the conditions under which they are preferable, , particularly as illustrated by U.S. federal courts. The same forces are at work throughout the economy, though with *unequal* influence.

The variety of measurement dimensions, the nature of the perils of motivating *better* performance with *greater* rewards, and why the perils are especially problematic in government and private nonprofit organizations, are further examined in other papers in this series. A central issue is the variety of ways to measure performance, and what I see as real outliers in both extremes the struggle over strong and weak rewards – ranging from "apple pickers" (who reap *strong* rewards for bringing in more apples), and federal judges (who receive essentially nothing for any type of measured performance).

In closing this paper, I make two brief observations: (1) I see two overarching themes: (a) all industries are the same, in major respects, and (b) all industries, whether dominated by for-profit firms, governments, or regulated private nonprofits, are different! This seeming contradiction offers insights to the advantages and disadvantages of tying rewards to "performance," as that is measured. Yet the goal of inducing, *better* performance continues, in all forms of enterprise.

Bibliography

Akerlof, George A. “The Market for "Lemons": Quality Uncertainty and the Market Mechanism,” *The Quarterly Journal of Economics* , 84 (3), August 1970: 488-500.

Alchian, Armen and Harold Demsetz. “Production, information costs, and economic organization.” *American Economic Review* 62, 1972: 777-795.

American Physical Society. “This Month in Physics History, February 1927: Heisenberg's Uncertainty Principle.” *APS News*, February 2008 (Volume 17, Number 2).

Beasley RP, “Development of Hepatitis B Vaccine*.” Journal of the American Medical Association.* 2009; 302(3): 322–324.

Bowers, Michael, Richard Hyde and Robert Wilson. “Office of the Governor: Special Investigators.” State of Georgia, June 30, 2011. http://www.11alive.com/data/pdf/CRCTInvestigationReport_Volume_1of3.pdf.

Business Insurance, “Volkswagen says diesel scandal has cost it $34.69B.” Business Insurance Magazine, March 17, 2020.) https://www.businessinsurance.com/article/20200317/NEWS06/912333571/Volkswagen-says-diesel-emissions-scandal-has-cost-it-$3469B-Frank-Witter-

Carlson, Joe. “Setting a standard: Illinois bill bases hospitals’ tax-exempt status on charity-care levels; plan could be model for other states.” *Modernhealthcare.com*, June 2, 2012. http://www.modernhealthcare.com/article/20120602/MAGAZINE/306029989

Cohen, Arthur and Florence Brawer. *The American Community College.* San Francisco: The Jossey-Bass Higher and Adult Education Series, 1996.

Creswell, Julie, Barry Meier, and Jo Craven McGinty. “Bills at Hospital in New Jersey Highest in U.S.,” *New York Times,* May 18, 2013: A1, B2.

Di Tella, Rafael. “Criminal Recidivism after Prison and Electronic Monitoring,” *Journal of Political Economy* 121 (1), February, 2013: 28-73.

Ewing, Jack. “vw

s Swift Journey From Scandal to Standout.” *New York Times,* March 20, 2021, B3.

Frost, Peter, “Legislation Defines Charity Care for Hospitals,”, Chicago Tribune, May 29, 2012. https://www.chicagotribune.com/business/ct-xpm-2012-05-29-ct-biz-0530-hospital-charity-care-20120530-story.html

Hallinen, Judith. “STEM education curricula,” *Encyclopedia Britannica*, June 2021. https://www.britannica.com/topic/STEM-education

Hansmann, Henry. “The Role of Nonprofit Enterprise,” *Yale Law Journal* 89 (5), April, 1980: 835-901.

Hartzell, Jay C., Christopher A. Parsons, and David L. Yermack. “Is a Higher Calling Enough? Incentive Compensation in the Church,” *Journal of Labor Economics*. 28(3), 2010.

Holmstrom, Bengt, and Paul Milgrom. “Multitask Principal Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design,” *J. Law, Economics, and Organization* 7: 24-52.

InfoPlease. “Life expectancy by Age, 1850-2011,” Sandbox Learning, 2012. https://www.infoplease.com/us/health-statistics/life-expectancy-age-1850-2011

Jones, Lori M. “Is It Bad Debt Or Charity Care? The Right Way To Measure Uncompensated Care.” November 11, 2016. *Health IT Outcomes.* November 17, 2020 at https://www.healthitoutcomes.com/doc/is-it-bad-debt-charity-care-measure-uncompensated-care-0001 .

Kreps, David M. *The Motivation Toolkit* (W.W. Norton & Company, New York, NY) 2018.

Kula, Witold. *Measures and Men.* Translated by Richafd Szreter. Princeton University Press, 1986: 15.

Lazear, Edward P. "Performance Pay and Productivity." *American Economic Review,* 90(5) 2000: 1346-1361.

Mahr, Joe and Robert McCoppin. “Study suggests racial mislabeling skews McHenry County sheriff data” Chicago Tribune, March 26,2011, https://www.chicagotribune.com/news/ct-met-mchenry-profiling-20110326-story.html .

National Vital Statistics System [NVSS]. “Provisional Life Expectancy Estimates,” p.2, Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/data/vsrr/VSRR10-508.pdf

Nierenberg, Amelia. “Student Reporters Expose Clusters at Colleges.” *New York Times,* November 5, 2020: A4.

Pérez-Peña, Richard. “Elite Colleges Differ on How They Aid Poor.” *New York Times,* July 31, 2013: A1, 12.

Politico. “Grassley tones down endowment threats.” *Policico*, September 10, 2008. http://www.politico.com/news/stories/0908/13340.html

Rosenthal, Elisabeth. “The Price for a Hip Replacement? Many Hospitals are Stumped, Research Shows,” *New York Times.* February 12, 2013: A16.

Salkind, Neil. *Encyclopedia of Research Design*, Sage Research Methods, 2010. https://methods.sagepub.com/reference/encyc-of-research-design/n175.xml

Sayare, Scott, “Rite of Passage for French Students Receives Poor Grade,” New York Times, June 27, 2013. https://www.nytimes.com/2013/06/28/world/europe/a-rite-of-passage-for-french-students-receives-a-poor-grade.html

Sears, Roebuck and Co. Catalogue, 1902 Edition, p.447.

Smalley, Andrew. “Higher Education Responses to Coronavirus (COVID-19).” National Conference of State Legislatures. https://www.ncsl.org/research/education/higher-education-responses-to-coronavirus-covid-19.aspx March 22, 2021

Smith, Adam. *An Inquiry into the Nature and Causes of the Wealth of Nations*, 1776.

Smith, Mitch. “Sen.Kirk calls for VA Hospital director to step down.” *Chicago Tribune,* March 30, 2017.

Stern, Ken. *With Charity for All* (New York: Doubleday), 2013.

Strauss, Valerie. “How and why convicted Atlanta teachers cheated on standardized tests.” Washington Post, April 1, 2015. https://www.washingtonpost.com/news/answer-sheet/wp/2015/04/01/how-and-why-convicted-atlanta-teachers-cheated-on-standardized-tests/

Supreme Court of the State of Illinois. Justices Karmeier, Fitzgerald, Thomas and Burke. Docket No. 107328. Provena Covenant Medical Center *et al.*, Appellants v. Illinois Department of Revenue *et al*., Appellees. Opinion filed March 18, 2010. June 4, 2013 at http://www.state.il.us/court/Opinions/SupremeCourt/2010/March/107328.pdf

Taylor, Art, Jacob Harold, and Ken Berger. “The Overhead Myth” letter at http://s5770.p9.sites.pressdns.com/wpcontent/uploads/2013/06/GS_OverheadMyth_Ltr_ONLINE.pdf June 17, 2013.

Taylor, Kate. “Parents Try Paying to Avoid Piper In a College Admissions Scandal.” *New York Times*, October 4, 2026: A1, 16.

Taylor, Kate. “Admissions Scandal Looms Over U.S.C. As New Year Begins.” *New York Times*, August 28, 2019: A14.

Thompson, Cheryl W. "Mystery over cases of award-winning D.C. Detective Milton Norris." February 18, 2012, May 9, 2013 at www.washingtonpost.com/investigations/confusion-over-numbers-for-award-winning-dc-detective-milton-norris/2012/02/10/gIQABxZUMR_story.html

Titmuss, Richard. *The Gift Relationship: From Human Blood to Social Policy* (London: Allen and Unwin), 1970.

U.S. Census Bureau. *CPS Historic time Series,* Table A-2, Percent of People 25 Years and Over Who Have Completed High School or College: Selected Years 1940 to 2020. https://www.census.gov/data/tables/time-series/demo/educational-attainment/cps-historical-time-series.html

U.S. Census Bureau, Education Attainment Tables, 2021. https://www.census.gov/topics/education/educational-attainment/data/tables.html

U.S. Department of Veterans Affairs, *Born of Controversy: The GI Bill of Rights,* 2006. https://www.va.gov/opa/publications/celebrate/gi-bill.pdf

U.S. Internal Revenue Service, “Excise Tax on Net Investment Income of Private Colleges and Universities,” Tax Cuts and Jobs Act – IRC Section 4968 Provision 13701. https://www.irs.gov/pub/newsroom/1-excise-tax-on-net-investment-income-colleges-4968-13701_508.pdf

U.S. Senate Committee on Finance. “Baucus, Grassley Write to 136 Colleges, Seek Details of Endowment Pay-outs, Student Aid.” January 24, 2008. http://www.finance.senate.gov/newsroom/ranking/release/?id=c7395dba-4033-48bb-b1ee-7b061f2e267e

Weisbrod, Burton A. “How Mixed Industries Exist: Modeling Output Choices in For-Profit, Public, Religious Nonprofit and Secular Nonprofit Organizations, with Application to Hospitals,” Working paper, Northwestern University, July 31, 2012.

Weisbrod, Burton A., Jeffrey Ballou, and Evelyn Asch. *Mission and Money: Understanding the University* (New York: Cambridge University Press). 2008.

Welch, H. Gilbert. ”Diagnosis: Insufficient Outrage,” *New York Times,* July 5, 2013: A17.

Winerip, Michael. “Ex-Schools Chief in Atlanta Is Indicted in Testing Scandal.” *New York Times*, March 29. 2013. June 4, 2013. http://www.nytimes.com/2013/03/30/us/former-school-chief-in-atlanta-indicted-in-cheating-scandal.html?hp&_r=0