Northwestern | INSTITUTE FOR POLICY RESEARCH

RESEARCH EXCELLENCE • POLICY IMPACT

IPR Working Paper Series

WP-25-5

Incentives in K-12 Education – The Perils of Not Just Measuring Performance by Standardized Tests, but Financially Rewarding It

Working Paper 2 in the Series: The Perils of Pay for Performance in Public Service Industries

Burton Weisbrod
Northwestern University and IPR

Version: February 5, 2025

Abstract

While the specific application of standardized performance testing for students has been debated over the years, the concept continues as a means of monitoring and encouraging better K-12 education. And yet, there remains a need to consider broader public policy issues of how incentives and undesired distortions occur, and what can be done to ameliorate them. In fact, standardized testing can and often does trigger a host of other actions and behaviors that may differ from intended outcomes. This paper examines the history of U.S. public policy regarding student educational performance measurement, its associated use, and resulting incentives. Through a series of case studies, it shows how strong rewards tied to simplistic testing-based performance measures can and do lead to undesired consequences. It discusses the economic concept of how and why measurement itself leads to changes in behaviors, and the rationale behind behaviors of ”gaming” testing systems to enable the appearance of better results. Finally, it points to the need for more sophisticated measurement of teaching performance with more nuanced use of “weaker” rather than “stronger” financial rewards as a way to better achieve desired results.

About the Series. This paper is the second 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 other specific industries including health care and higher education.

Acknowledgements. The research and discussions of performance testing for K-12 education was assisted by many people over the course of a decade of work. For this paper, I am particularly grateful for important editorial assistance and discussion refinement provided by Northwestern University's Dr. Evelyn Asch, now retired, and by Michael Gnat, a free-lance copy editor. I also thank the Northwestern University students who served as research assistants, specifically William Russell, Grant Johnson, and Camille Liu.

Introduction: A Historical Background on Standardized Testing

The *use* of standardized testing at public schools has evolved over time, but the *incentives* created by it endure, along with broader and often unintended consequences that continue today. This paper highlights these issues and suggests a way to address them.

When the No Child Left Behind (NCLB) Act became U.S. law in 2002, it ushered in an unprecedented era of massive, standardized testing in K-12 public schools across the country. In addition to mandating testing in those schools to qualify them for federal support, the Act linked the monetary rewards to student test scores; better performance by students meant greater rewards to their schools and administrators. The goal was to invigorate stakeholder incentives -- for schools, students, teachers, and administrators, all to elevate education system performance.

While it was subsequently replaced by a later federal law, NCLB represented the start of a longer-standing federal effort to induce "better" state-administered K-12 education. But it was by no means entirely new. Far from it. More than two centuries earlier, in 1789, the fledgling United States of America had adopted a constitution that left public education as the responsibility of each state, not of the federal government. NCLB advanced an era of demanding quantitative evidence of schools' accomplishments as a precondition of their receiving performance-based subsidies from the national government.

And much more; it also set the stage for future federal subsidies for other public service industries such as hospitals, police, charities, and museums. By today's 21st century performance-based incentive standards, linking federal financial rewards to measured performance on standardized tests, in English and mathematics, and later in science, throughout the economy, seems wise if not long overdue. But there were also unintended side-effects of the changing incentives, as stakeholders saw new opportunities to advance their own self-interests, even adopting fraudulent and otherwise illegal methods that were not attractive with weaker incentives.

Only seven years after introduction of the NCLB legislation with its stronger incentives for measurable performance, federal support emerged for K-12 schools based on their *students'* achievements on standardized tests in math, English, and, later, in science. Multiple quantitative measures of students' and schools' performance were adopted, and federal grants to states for their K-12 schooling were tied to those subjects and measures in the 2009 "Race to the Top" legislation.

Dramatic, yes, but the idea of connecting student scores on particular standardized tests, to federal grants for K-12 schools was not novel; 175 years earlier it was already generating ties between *students'* performances on standardized written tests, and *schools'* receipt of monetary rewards from governments.

Written tests to evaluate student and school performance had first appeared in the United States as early as 1845, when Horace Mann and members of the Boston School Committee developed written exams to gauge student performance in public schools, replacing oral exams. The often-heated debates between traditional *Master Teachers*, who controlled the *oral* exams, and Mann and other supporters of standardized, *written,* exams, became front-page news and led to the first serious consideration of written tests as useful pedagogic tools and measures of student achievement. Additional links,

specifically to teacher pay, came later, although even in Mann's days and earlier, "Failure at school had its price, from the shame of repeating grades to the failure to progress academically."[1]

Under NCLB, entire schools were sometimes sanctioned for failing to achieve "Adequate Yearly Progress" as measured by standardized tests. And pay-for-performance expanded far beyond rewarding or penalizing student excellence; it affected teacher salaries, administrator promotions, and pressures on students' parents and on state legislators and governors, who recognized that low-performing students were causing loss of federal grants. That could be mitigated ina variety of ways through gaming to elevate average test scores. Some potential examples included increased home-schooling, delayed starting of kindergarten (red-shirting), so that throughout the entire K-12 schooling years, students would be older, wiser, and stronger than their classmates, and so would be likely to perform better; additionally, parental opting-out of their children from standardized testing, and, of course, blatant cheating on exam grading raised student scores and cut federal grants to the schools based on "need."

The NCLB Act was barely a dozen years old when its performance measures and rewards were replaced by a new Federal law that moved more responsibilities to the states. Focusing on schooling, the "Every Student Succeeds Act" (ESSA) of 2015 permitted flexibility for each state to establish its own K-12 schooling goals and methods for achieving them, though still requiring each state to institute a plan for testing and measurement of goal achievement. That left the federal government with a continued role for financial support and incentive grants. So even with this added flexibility, the concept of school testing and its associated incentives continued.

A Pay-for-Performance incentive structure encourages particular actions and discourages others -- by teachers, administrators, employers, regulators, consumers, or other stakeholders – and shows little if any sign of diminishing throughout the economy. Rather, it reflects an expanding perspective; increase or decrease the rewards or penalties to any stakeholders, and there will be responses. Whatever the industry, and whether it is owned and controlled by a government agency, a private nonprofit organization, or a for-profit firm, incentives will differ, often through the tax and subsidy systems, and performance will be affected. The obstacles to defining, measuring, and valuing "better" and "worse" performance in much of the economy hinder development of mechanisms for measuring performance and establishing efficient incentives.

So, getting incentives "right" is the challenge -- how to measure and implement rewards and penalties that neither systematically exceed nor fall short of the value of additional output. In settings where performance and its rewards can be easily gamed, to attract excessive compensation, the result is that the forms of performance that are most easily adopted are those that are also easiest to game and reap the rewards. Measured, not true, performance, is maximized, which often differ.

Why has all this happened—in little more than two decades? Why particularly in K-12 schools? And were the 2002 NCLB and 2015 ESSA Acts harbingers of unexpected side-effects of higher stakes for measured performance throughout the economy but particularly in the public and private nonprofit sectors, and not only in schools?

[1] Reese, *Testing Wars in the Public Schools,* 231

I introduced the concept of Pay-for-Performance (P4P) for government and nonprofit programs in paper 1 of this series,[2] as a term applicable for any public service industry that derives financial or other related incentives and rewards from performance outcomes regarding quantity or quality of service provided. Understanding what P4P *actually* accomplishes in schools, and not merely what it is intended to accomplish, is critical for efficient public policymaking. The central question is whether the motivational foundation of a private enterprise economy—rewarding *observable, measurable,* performance— also makes sense for schools.

To answer this requires identifying the limits (and limits there are) of a well-specified P4P methodology to stimulate efficiency in any industry. And in schools, the differential roles of rigidly defined tests for measuring performance by, and rewards to, teachers of diverse subjects, to students of diverse backgrounds, abilities, motivations, and parental support, as well as to gauge administrators' performances in diverse settings.

The successes and failures of strong rewards for measured K-12 performance turn out to provide valuable lessons ranging far beyond schools. But what those lessons are is not self-evident. They include, though, the inherent dangers of strong rewards, despite their superficial appeal. The dangers reflect incentives that are not merely imperfect but are systemically *incomplete* and *biased* -- encouraging *over*statements of *performance* when that would increase stakeholder compensation, and encouraging *under*statements of stakeholders' shortcomings when that would downplay their weak performance and diminish the penalties for weak performance In either case, the distorted incentives drive a wedge between rewards and true accomplishments.

The dangers that strong rewards will be "gamed" and incentives exaggerated are clear enough in some industries that strong incentives are in fact little used; in other industries, including K-12 schooling, there is broad, though fading, consensus, whether warranted or not, that the dangers of strong rewards are not severe compared to their benefits. Evidence is scarce, though, and anecdotal. But in the longer run, we can expect to see the use of strong rewards withering away as gaming of incentives becomes more widespread, sophisticated, costly to prevent or detect, and subject to civil and criminal law enforcement.

With respect to the future of links between *measured* performance and its rewards and penalties—often thought of as "merit pay" -- we can expect a policy reversal ahead – onto a path of *weaker* rewards for forms of performance that are easily and accurately *measured and valued.* This shift away from *increased* emphasis on measuring performance and rewarding it, to *decreasing* measurement and rewarding of performance, is surprising but also thought-provoking: Strong and weak rewards for better performance are certainly different, though also very much alike, in their pursuit of ways to tailor incentives so as to facilitate exchanges between buyers and sellers. How the diverse segments of an economy can be both the same and different is a puzzle from which there is much to be learned as we struggle to understand it and show how it will re-shape our future.

In the case of K-12 schooling the evolution of professional judgement on whether and how "better performance" should be measured and rewarded is already evident over the past two decades. Education historian and former U.S. Assistant Secretary of Education, Diane Ravitch, was an early and strong proponent of standardized testing in public schools, and of matching of those test results with

[2] Weisbrod, "Why Strong Performance Rewards in Government and Nonprofit Programs Don't Work"

federal rewards under the No Child Left Behind Act of 2002. But Ravitch dramatically reversed her supportive view in her 2010 book, *The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education;* she was highly influential in reversing the rising tide of support for both national standardized testing and its linking of school performance on the tests to federal government financial rewards.[3]

When Are Strong or Weak Incentives Efficient, and What Does That Mean for Schools?

There is no "one size fits all" reward structure; much depends on the measurement, as well s the valuation, of performance. When performance is measured incompletely and so with statistical bias, as is virtually inevitable, strong rewards are likely to be inefficient, driving incentives in counterproductive directions that often achieve the opposite of the intended goal. In short, under those conditions, the path to greater efficiency may well be through adoption of *weak* rewards.

A Lesson on Strong Rewards for Better Performance: in Congress and in Schools

To set the stage for understanding why strong rewards can be *in*efficient for schools, consider the challenge of rewarding performance in a quite different realm -- for members of Congress. In May of 2014 the 113th Congress was termed the worst-performing, least "productive," in history, based on a simple, low cost, measure of performance, the number of bills it passed. The poor performance claim looked even more dire if the number of bills passed was measured by excluding bills that merely changed the names of individual post offices, commemorated anniversaries, or announced other ceremonial matters.[4] Yet despite the ostensibly abysmal performance record, there was no call to cut Congressional salaries nor in any other way to tie salaries to the number of bills passed nor to any other measure of performance. Why not?

To do so, to align rewards with the number of bills passed, which is easily measured but also easily gamed, would turn an *indicator* of accomplishment into a basis for *rewards*, and in the process would distort Congressional incentives. Why pass one bill covering numerous ceremonial actions, when it could be broken into many less-comprehensive bills doing the same things, but in the process passing more bills, increasing legislators' measured performance and their compensation? Incentives matter.

It would be foolhardy and inefficient to incentivize Congress to disaggregate complex legislation into multiple bills in order to simply pass more laws and augment compensation. Such a reward system would be so obviously inefficient that it was not even considered. Not even by critics of the "do-nothing" Congress. True, it could have generated increased *measured* performance. Also true, if Congressional compensation were tied to the number of bills passed -- implying that all are equally "important" -- lawmakers' incentives would change, and more bills would be passed.

In K-12 education the perils of rewarding schools and teachers for having more of their students *pass* standardized tests are less evident, but the leap is not as large as it may seem. The dangers for schooling are essentially the same type as for Congress – exaggerate not *real* performance, but its poorly-*measured* form, to reap greater compensation.

[3] Ravitch, "The Death and Life"

[4] Blow, The Do-Even-Less Congress"

What is True Performance in K-12 Schools, and How to Measure It?

The NCLB Act had elevated two easily obtained measures of educational "success"-- the percentages of a school's students who passed the state's standardized, generally multiple-choice, tests in two particular subjects -- reading and mathematics, and, later, in science —to unprecedentedly high levels of rewards. The ESSA Act, which still applies today, largely maintained that objective even as it provided greater state level flexibility. As a consequence, teachers, schools, principals, and other stakeholders were often rewarded for higher rates of students passing the "objective" tests, based on each state's own standards of what constituted a "passing" grade, and how a particular student's answers should be scored and rewarded, particularly in light of the fact that "better" answers brought rewards to the school and its personnel, though at costs that were borne largely if not entirely by outside grants from the federal government.

Still, students' scores on the reading and math exams, and in some states, in science, brought unintended side-effects. Some teachers were rewarded with job tenure, were promoted (or discharged), and some schools received federal education grants, [5] depending on their students' test scores, so there were clear incentives for teachers and administrators to raise *reported* test scores by any number of means—encompassing such forms of gaming as encouraging low-performing students to quit school, be home-schooled and so be excluded from school performance statistics, or cheat in the grading process, .

How a school's performance is *measured* is thus increasingly salient when it is linked to rewards and penalties. Efforts to reward teachers for their "value-added" to students' lifetime earnings seem appealing, but deriving unbiased estimates of the portion of students' lifetime earnings that is attributable to particular schools and their teachers is enormously demanding. Even estimating the effect of a particular teacher on a particular student's English or math test score, which may or may not affect future earnings, is exceedingly difficult in light of students' varied learning abilities, motivations, and home environments.[6]

Many of those forces are largely beyond the control of teachers and school administrators, including family poverty, student health and absenteeism, and parental support. Other variables affecting learning in school include teachers' and principals' decisions to allocate more classroom time to math, reading, and, later, to science when it became subject to a standardized test, and less to social studies, philosophy, art, physical education, and other untested subjects.

Consider student absenteeism, which often contributes to poor test scores. Many of the underlying problems and learning handicaps it reflects are associated with poverty, which may be concentrated in particular neighborhoods and regions. In Florida, for instance, chronic school absenteeism—students missing at least 10 percent of class days in a year—has been very concentrated: 52 percent of chronically absent students have been found to be concentrated in 15 percent of the Florida public schools, a density that surely handicaps teaching and learning.

In Maryland the link of poverty to poor academic performance, apart from the quality of instruction, is suggested by the fact that 31 percent of low-income high school students who were poor enough to be eligible for the federal lunch program were chronically absent from school, compared to

[5] Rich, Home Schooling: More Pupils, Less Regulation"
[6] Hanushek and Rivkin, Value-Added Measures of Teacher Quality"

only 12 percent of students above that income threshold.[7] Low family income, affecting school absenteeism, and inadequate diets are other impediments to student learning, further confounding efforts to reward schools, teachers, and administrators by relying on their students' standardized test scores. These are daunting obstacles.

Gaming and the Unintended Consequences of Highly Rewarded Test Scores

A multiplicity of performance variables and their interactive effects on test scores provide ample opportunities for schools, teachers, administrators, and parents, to game a reward system that hinges on those scores. Incentives for gaming are at the root of the public policy controversy over how test scores should be used to assess t*eacher* performance and compensate it, and they underlie the very real perils of pay for performance at K-12 schools. In fact, the greater the rewards to teachers or others when student test scores rise, the more perilous P4P becomes—the greater the probability that *true* learning goals will not have been achieved, even as *measured* goals are. On the other hand, there is good reason to believe that when rewards for higher test-scores are negligible, they are likely to have little or no effect. When better test scores are rewarded generously, the rewards for stakeholders will stimulate development and application of new forms of gaming.

Stronger rewards encourage greater efforts by stakeholders to reap them—whether by genuinely improving *learning* or by only improving *test scores*, proxies for learning. Ideally, the two paths would be the same; realistically, strategic manipulations of the testing and grading processes can be gamed, improving scores *without* improving learning. Moreover, it is costly to distinguish real improvements from gamed ones. When teachers or administrators manipulate results, they typically understand that what they are doing is *not* the true goal of the program, and so they have an incentive to hide their behavior in markets with informational asymmetry between producers and consumers, as emphasized in paper 1 of this "Perils of Performance" series.[8] Greater rewards for improved test scores increase incentives to mask test scores by means that are costly to detect.

The crucial efficiency issue, as Adam Smith explained nearly 250 years ago, in 1776, in his *Wealth of Nations* book*,* was whether the pursuit of a stakeholder's *private* self-interest -- which, in schools today, would emphasize teachers and principals -- or also *collective*, *social*, benefits to a wider population. To advance social efficiency, incentives should not simply advance the *private* interests of particular stakeholders. If, for example, incentives rewarded *only* higher test scores in specific subject areas, higher scores in those subjects *will* result—but not necessarily because of greater student learning or more effective instruction, but via gaming. If better test scores in English, math, and science are rewarded more generously, those scores will rise, as class time and homework are increasingly shifted to those subjects, notwithstanding the negative effects on other subjects.

A Gaming Matrix: Legality and Stakeholders

Gaming takes innumerable forms in education and throughout the economy, but they fall usefully into two broad types. One involves its legality, the other, the strategic position of organization administrators to impose rewards and penalties. A matrix of who engages in gaming and what forms it takes, highlights the complexity of detecting and preventing gaming.

[7] Cardinali, "How to Get Kids to Class"

[8] Weisbrod, "Why Strong Performance Rewards in Government and Nonprofit Programs Don't Work"

*G*aming can be illegal, legal, or borderline. Even if legal, though, it may not be socially efficient, since gaming strategies vary in their cost to implement, on one hand, and to detect, on the other, as well as in the severity of their distortionary effects. Though the specific forms of gaming vary across the industries discussed in this and other papers in the "Perils of Performance" series, it is noteworthy that the most egregious gaming—the blatantly illegal forms—are but the tip of the proverbial iceberg: legal and borderline forms of gaming dominate, even if they attract fewer eye-catching headlines.

Who Engages in Gaming the K-12 Incentive System and its Matrix?

Any participant in a P4P process has the potential ability to game its reward structure. For K-12 schools, there are four major participants—parents, teachers, school administrators, and state or federal governments—but others such as students (especially at the high school level) and purveyors of capital equipment and textbooks can also take advantage of reward opportunities. The multiple combinations of legal status and types of stakeholders, shown in Figure 1 below, highlight the variety of forms gaming may take in K-12 schools. The three levels of legality to be discussed, and the five types of stakeholders to be discussed, imply fifteen potential types of gaming mechanisms—even more than that when we recognize that, for example, parents and teachers have many legal instruments available to influence test scores. Strong P4P rewards can be manipulated, and whatever the stakeholder's intent, the result can undermine the intended performance goal. In the process, regulatory attempts to make schools more accountable may encourage the opposite of what is intended.

A Matrix of Stakeholders

Every stakeholder, by being involved in a production process, has an incentive to act in ways likely to garner the most rewards—to achieve the highest level of rewarded performance. Every stakeholder also has choices of how to do that—particularly by choosing among various legal, illegal, and borderline legality methods. While the particular stakeholders and their options differ among program areas, the value of the matrix is that it calls attention to the *potential* actors in the gaming process and the variety of forms their gaming can take. Whether the focus is on schooling, health care, policing, charities, the judiciary, or any other nonprofit or governmental activity, identifying the stakeholders is the first step in the process of knowing where to look for evidence of gaming any P4P reward system. The schematic below portrays the principal parties gaming reward systems in the three programs examined in this paper.

Figure 1. *A Performance Gaming Matrix: Legality, Stakeholders, Applications to K-12 Schools*

A Gaming Menu	*Stakeholders*					
	Managers (1)	Teachers (2)	Parents (3)	Governments (4)	Students (5)	Other (5)
Illegal Gaming	●	●				
Legal Gaming		●●	#			
Borderline Legality		●●●	##			

● *Cheating on grading student test scores under the No Child Left Behind Act (Atlanta case)*
●● *Re-grading to raise exam scores from below "passing" to passing (New York State case)*
●●● *Selective expulsions of low-performing students to raise school test averages (Florida case)*
 # Redshirting: delay of school enrollment to gain competitive advantage over younger students
Home schooling: a device for evading (opting-out of) mandatory standardized tests

The cases sketched in Figure 1 are examples intended only to illustrate the wide array of mechanisms through which program objectives may be gamed and evaded. More details are presented later.

In the public schools in Atlanta and Philadelphia, the elevated rewards to teachers and administrators for their performance excellence—measured by higher standardized test scores — proved not only *un*productive but *counter*productive; P4P brought negative results. Bonuses to teachers and school administrators for meeting or exceeding testing standards encouraged educators to cheat, and some succumbed, resulting in criminal charges and imprisonment. But whatever the legal status, the mechanisms involved gaming – use of mechanisms for circumventing intended restrictions or mandates.

The experiences of these and other cities highlight two conclusions:

1. Higher *test scores* are not equivalent to increased *learning*, and
2. We know how to measure test scores, but not how to measure "learning."

In many cities, rewards to teachers and administrators for higher test scores—through bonuses, promotions, and acclaim--combine with social pressure from colleagues to improve test scores, but that approach is a devilish recipe for gaming. From 2009 to 2011, for example, in some Philadelphia schools, "administrators were giving correct answers to teachers who passed them on to students. In other cases, principals took completed exams home at night and doctored the answer sheets."[9] In Atlanta schools, teachers and administrators, with support from the Superintendent of Schools, held after-school get-togethers to "improve" test scores by erasing students' wrong answers and correcting them. Of course, test scores increased.

That was one path to better performance, as measured—a path using test-score gaming manipulations that had been claimed to be common across the country and to "have been confirmed in at least 39 states and Washington, D.C." [10]

Another incentive factor affecting teachers was pressure from parents to help students get through a state mandated test score requirement for high school graduation. With support from teachers' unions and some parent groups, that number of states with such a requirement has fallen from 27 a decade ago to only 7 as of the end of 2024.[11] And while each state now controls the content and use of its standardized testing, the federal law requiring standardized testing and its interpretation in judging school and district-wide education performance still endures for all 50 states. Thus, it is worthwhile to further examine how some school districts have gamed the system over time. Past cases that received media attention illustrate this problem.

Gaming and Cheating in Atlanta Public Schools

In 1999, three years before the NCLB financial incentives went into effect, poor performance had already become an issue in the Atlanta school system, which included more than a hundred schools and

[9] Rich and Hurdle, "Erased Answers on Tests in Philadelphia

[10] Fausset, "Trial Opens in Atlanta School Cheating Scandal"

[11] LeBlanc," More states ditch exams as high school grad requirements"

approximately fifty thousand students. To address those schools' dismal education situation a new and highly respected replacement, Dr. Beverly Hall, was hired that year as the new Atlanta Superintendent of Schools. She was aware of the school system's reputation for poor performance, including treating homework like a joke and doing little to halt drug dealing in front of some schools.[12]

Hall's innovations involved more than illicit *measurements* of student performance, and then holding teachers accountable for improving test performance compared with the prior school year's results.[13] Hall was under great pressure to raise students' performance as determined annually by the federal legislation's statistical measures, allowable grading methods, and efforts to enforce rules against schools' cheating or otherwise misrepresenting students' true performance.

Seeing strong incentives as the route to success, Hall used both rewards and penalties to motivate schoolteachers and other employees: there were bonuses of up to $2,000 per employee when a school's student test scores met "improvement" targets, and principals whose schools did not meet their targets within three years were fired. Her quantitative targets included a minimum 3 percent annual increase in the number of students meeting or exceeding Georgia's test-score standards. Overall in the course of Hall's eleven-year tenure, 90 percent of Atlanta's school principals were replaced.[14]

Teachers also faced *negative* monetary reinforcement. Even if teachers whose student test scores did not increase materially -- however that was achieved -- they could hold onto their jobs, but they could still be hurt financially. If federal funding for Atlanta schools was cut, as was likely if standards were not met, their salaries and jobs could suffer.

The strong incentives worked. Test scores in the Atlanta schools mounted rapidly. Between 2002 and 2009, eighth graders' reading scores on standardized tests jumped fourteen points, a greater leap than for any other urban-area school district.[15]

Not only did teachers and principals reap financial rewards, but student scholarships also rose from $9 million to $129 million, and a billion dollars was spent on school buildings. Hall, too, profited: in 2009 alone, she received a $78,000 bonus,[16] almost 30 percent of her base salary of more than $270,000,[17] and between 1999 and 2009 her performance-based bonuses totaled more than $580,000.[18] She was also rewarded by being named National Superintendent of the Year by the American Association of School Administrators (AASA) for her "leadership, communication, professionalism, and community involvement."[19] AASA Executive Director, Dan Domenech emphasized test-score improvement, noting that "Hall has accomplished significant gains in student achievement. She has demonstrated a commitment to setting high standards for students and school personnel."[20]

---

[12] Aviv, "Wrong Answer", 57
[13] Bowers, et al., *Special Investigation into Test Tampering,* 351
[14] Aviv, "Wrong Answer", 57
[15] Koebler, "Educators Implicated in Atlanta Cheating Scandal"
[16] Beasley, "Jury Chosen in Cheating Trial"
[17] Severson. A Scandal of Cheating, and a Fall from Grace"
[18] Baird-Remba, "Former 'Superintendent of the Year"
[19] AASA, "South Carolina, Texas Educators "
[20] Aramark , National Superintendent of the Year."

Even as kudos and cash poured in for Hall and others, skepticism of the spectacular test-score improvements was growing. Investigations found that Hall had been involved in one of the nation's largest school test-cheating scandals ever.[21]

The test-score successes were clear, but the probabilities that true learning would jump so sharply seemed very low, and so the startling performance growth began to attract suspicion. Only a few months after Hall received the 2009 bonus and national acclaim, the *Atlanta Journal-Constitution* newspaper published a series of reports noting the untrustworthy performance gains at Atlanta public schools, and then-Governor Sonny Perdue launched a fifty-person investigation of 56 of Atlanta's 104 public schools, and the interviewing of more than 2,100 individuals, including staff and students at those schools where performance statistics were particularly questionable.[22] By 2011, irregularities were uncovered in 44 of the schools, and 82 of the 178 implicated educators later confessed to cheating, largely by erasing and correcting student answers.[23]

In March 2013 a Fulton County, Georgia grand jury indicted Hall and thirty-four other school employees -- four of her executive administrators, six principals, two assistant principals, fourteen teachers, a school secretary, a school improvement specialist, and six testing coordinators.[24] There were indictments on sixty-five counts including criminal conspiracy to boost test scores, making false statements during the investigation, and intimidating witnesses in an effort to hinder the investigation.[25] Since then, at least one defendant died, prosecutors agreed to plea bargains with twenty-one others, while Hall's own trial was delayed indefinitely because of illness—she was undergoing treatment for stage-four breast cancer.[26] She denied, however, knowledge of any cheating in the Atlanta schools,[27] and she died in 2015.

On September 22, 2014, after approximately six hundred potential jurors were questioned individually over a period of six weeks, the jury was chosen for the trial of the remaining twelve former Atlanta educators who had since resigned or been fired.[28] When the trial began a week later, the scope of the alleged violations was presented by prosecutor Fani Willis, who said that students would attest that teachers had given them the correct answers to standardized tests; other teachers would admit that they had worked with the defendants to alter test answers; and administrators testified about conspiring to boost student test scores.[29] The defendants faced up to thirty-five years in prison on felony charges.[30]

In Atlanta, the strong P4P rewards and penalties were clearly failures, if not disasters. (There remained the possibility, though, that some *other* teachers and administrators did *not* game the system but truly did improve the education of their students.) Responding to incentives by cheating — taking

[21] Jarvie, "Atlanta School Cheating Trial"
[22] Copeland, "School Cheating Scandal"; Jarvie, Atlanta School Cheating Trial"
[23] Layton, "GAO: 40 States Have Suspected Cheating on K-12 Tests"
[24] Carter, "Grand Jury Indicts 35"
[25] Jarvie, "Atlanta School Cheating Trial"; Carter, Grand Jury Indicts 35"
[26] Jarvie, "Atlanta School Cheating Trial"
[27] McWhirter and Banchero, "Ex-Head of Atlanta Schools Indicted".
[28] Beasley, "Jury Chosen in Cheating Trial"
[29] Bloom and Rankin, "Heard Most on Day 1 of APS Trial?"
[30] Jarvie, Atlanta School Cheating Trial"

advantage of the high costs of being detected and seriously penalized —was appealingly simple. It was also inexpensive compared to actually improving the educational system.

Cheating—fraud—has been alleged to have occurred in schools in thirty-nine states in recent years, and in a variety of forms.[31] A federal report noted that "officials in 33 states confirmed at least one instance of cheating" during the school years 2010–11 and 2011–12.[32] In Atlanta, educators even went so far as to correct student answers *after* exams; elsewhere, other illegal but effective ways to "improve" test scores sufficed.

More Cheating: in Philadelphia, Washington, D.C., and Other Public Schools

In a Philadelphia elementary school between 2006 and 2011 the principal and four teachers allegedly boosted student standardized test scores through another illegal method; students were given the correct answers. A grand jury investigation found that some students had been instructed to record their test answers on scrap paper instead of the test booklets so that teachers could check the answers in advance.[33] Others were guided *during* the test; the Pennsylvania Education Department found that approximately half of all third graders at a school had five or more incorrect answers on a test erased and corrected, while 15 percent had more than ten such changes of answers. The educators were criminally charged with tampering, forgery, and criminal conspiracy; two were also charged with perjury and felony racketeering.[34]

After the cheating had been exposed, the schoolwide student *proficiency* on reported test scores quickly plummeted. In math it dropped from 41 percent for the year 2010–11 to 22 percent for the next year; in reading, from 40 percent to 20 percent.[35] For third-grade students the 63 percent "passing" the math test collapsed to 30 percent after the cheating exposure, and on the reading test the drop was from 60 percent to 27 percent. Cheating had been massive. [36]

Following those scandals, more than fifty other Philadelphia schools were investigated, and sixty-nine current and former employees were found to have acted not just improperly but perhaps illegally. At least three school principals were fired, and at least twelve educators were otherwise disciplined.[37]

School performance, as measured by test scores, reflected the new incentives, as penalties for cheating in the Philadelphia Public School District were levied, enforced, and publicized. In the 2011–12 school year, after the districtwide scandal, the number of Philadelphia schools reporting achievement of the state's Adequate Yearly Progress (AYP) standard dropped precipitously, by 70 percent in one year, to 33 percent.[38]

Improvements in *reported* student performance in other cities and states were also "too good to be true." In Massachusetts a principal resigned after the stellar performance of the school's students—the percentage who passed a state test and so were reported to be "proficient" in the subject— had

[31] Beasley, "Jury Chosen in Cheating Trial"
[32] Layton, "GAO: 40 States Have Suspected Cheating", 3
[33] Lattanzio, "Philadelphia Teachers, Principal Charged"
[34] CBS/AP, "Philadelphia Principal, 4 Teachers Charged"
[35] Pennsylvania Department of Education, "2010–2011 PSSA and AYP Results".
[36] Lattanzio, "Philadelphia Teachers, Principal Charged"
[37] Graham, et al, "Charges in 'Culture of Cheating"
[38] DeNardo, "With New Anti-Cheating Procedures"

more than quadrupled but then suddenly fell then from 17 percent to 76 percent of the class.[39] A seeming miracle!

Cheating by Teachers and Administrators in Washington, D.C. Schools

Public schools in the U.S. capital reported in 2017 that their overall graduation rate for that year was the highest in the history of the school system; about 73 percent of all students graduated "on time" – within four years. The school system had struggled for years to increase that graduation rate above 50 percent. The 73 percent graduation rate was four points greater than it was the previous year, and 20 points above its 2011 level. The Mayor of Washington, DC proudly described its school system as the "fastest improving urban school district in the country."

But it was all false. A 2018 report by the Office of the State Superintendent of Education showed that more than one-third of the diplomas awarded to students in 2017 were not earned. The report found that 937 out of 2,758 graduates of D.C. public schools did not actually meet the minimum attendance requirements needed for graduation. Teachers even admitted to falsely recording students as "present." [40] Washington, DC was one of the many public school systems found guilty of widespread cheating. Others included Chicago, Memphis, Los Angeles, and Columbus, Ohio, in addition to others noted above.

The perpetrators of these scandals weren't students, they were school administrators and teachers. Many admitted to falsifying records on standardized test scores and on students meeting other graduation requirements.

In response, some teachers were fired and stripped of their licenses to teach again. Some were sent to jail. Antwan Wilson, District of Columbia schools Chancellor, resigned after it was revealed he used his position to get his daughter into a preferred school.

The real culprit in these cheating scandals was pressure from federal, state, and local governments to achieve better performance. If mandated standards were not met, school systems could lose funding.[41]

Responses to Evidence of Cheating

Cities and states throughout the country have responded to opportunities and incentives to "improve" their measured performance, frequently through various forms of cheating. At least eight states, including Florida, Idaho, Indiana, Kentucky, Massachusetts, and Mississippi have implemented external anti-cheating mechanisms such as hiring anti-cheating consultants, using software and statistical analyses to detect cheating, using outside proctors to oversee standardized testing, setting up anonymous tip lines for reporting cheating, and switching from paper to electronic exams. Many states contracted with one large private firm, Caveon Test Security, in a growing private industry providing cheating-prevention services for schools. As the incentive to cheat on test scores grew, an education specialist, Gregory Cizek of the University of North Carolina, summarized a memorable lesson: "Nobody wants to be Atlanta."[42]

---

[39] FairTest, "NCLB Boosts Temptation to Cheat"
[40] Alvarez and Marsal, *Audit and Investigation*
[41] Howard University, "Why Public School Teachers"
[42] McWhirter and Porter, "Measures to Detect Cheating"

Between Legal and Illegal: "Borderline" Gaming of New York State Regents Exams

In New York State, all public high school students have to pass a set of subject based Regents exams to graduate from high school, and those with high scores get a coveted Regents diploma. That raises an issue of what is defined as passing, and what is legal to affect test outcomes.

Legality is not always clear. When judgments are required for grading essay exams—when there is no well-defined "correct" answer—incentives will affect the way judgments are made on grading. Some elements of a high-quality education involve students learning identifiable material that can be well-tested objectively as being *right* or *wrong*. Other elements, such as the "quality" of written expression of ideas, are not so easily tested, requiring discretionary judgments by exam graders. How those judgments are made affects individual student exam scores, class averages, passing rates, Annual Yearly Progress, and other statistics that bring rewards to teachers, students, and school administrators for their "good performance."

How judgments on grading are made *could* have little or even no effect on class grade averages and the percentage of students demonstrating "proficiency"—that is, receiving a grade deemed "passing," That is what is expected if a grader's judgments do not depend on the assessment of the consequences of exercising judgment in one or another direction and degree. If a grader's judgments are essentially random—sometimes being a little generous, other times a little stingy—*average* statistics on performance will be unaffected. But those are big "ifs."

This is not what to expect when performance success is measured by the percentage of students that reach or "pass" a specific but arbitrary threshold, and when only then is greater success rewarded. Whatever forms the rewards take—money, prestige, or even the grader's personal satisfaction from seeing more students pass—the result is the same: the prospect of greater rewards brings systematic grading biases, not random variation.

When a student who scores 65 on a standardized exam "passes," while one who scores 64 "fails," even that single point can have major implications for the student, parents, teachers, school principals, and other administrators. All share an incentive to use their influence and discretion to "nudge" marginal students up to the "passing" threshold. Those incentives can be seriously distorting, systematically increasing the percentage of a class or a school that is reportedly "proficient." Over time, the upward bias can grow as the rewards for passing grades produce a declining failure rate, a form of grade inflation.

It is not easy to determine how teachers grading test answers exercise their discretionary power to increase the number of students with passing scores -- generating coveted Regents Degrees, for example, which are valuable in gaining admission to selective colleges. Yet the test scores are visible, and they do not necessarily involve any "cheating" or illegal activity at all. Friendly, sympathetic, supportive judgements can suffice to elevate test grades. To illustrate: evidence has surfaced for the New York State Regents Exams in English, math, global history, U.S. history, and science, all of which must be passed for a student to receive a valued Regents graduation degree.

Stakes are high: "For the 2009 Regents exam in English, for instance, students were more than five times as likely to get a test score of 65—the minimum *passing* grade—than they were to score one point lower, and fail. In the U.S. History and Government Regents exams, students were 14 times more likely

to get a 65 than one point lower."[43] Was this bunching at the threshold of passing simply chance? Unlikely.

Clustering at the pass–fail cutoff appears to result from a confluence of forces, all affecting incentives and opportunities for gaming—bringing an upward bias of the grading system at the pass-fail boundary. In the case of New York State, education officials ruled that teachers *re-grade* a Regents test if a student barely *fails*, in order to check for grading "errors." The cautious policy is understandable, as a signal to graders that a small score increase could bring a large and potentially life-changing effect, especially since the policy (subsequently changed) had been for teachers to re-grade their own students' tests.[44]

With judgments required of teachers grading essay questions, and, in math, where student explanations of how they arrived at their solutions, and teachers' conflicts of interest are clear, teachers cannot serve simultaneously as "multi" agents for all these stakeholders, the State, and individual students. The effect of a test re-grading decision on a particular student's exam makes it likely that strict adherence to State standards for determining who passes will not prevail. And unless a teacher holds animosity for a particular student, the expected result is that re-graders, whether for their own students or not, establishes an incentive to add a point or so to raise a test score to the passing level. Bunching of scores just above the boundary is predictable.

And the bunching of test scores was found. Independent analysts found statistical evidence of bunching at the 65 passing level on the Regents Exams. Judging from the smooth cumulative distribution of scores throughout most of the range from zero to 100, the number of grades just a point or two below the passing threshold of 65 was deemed substantially *under*stated, and the number of grades at or slightly above 65 was *over*stated. The conclusion: 3–5 percent of all NY State students statewide who "passed" the five Regents exams in June 2009 "actually failed."[45]

But whether the bunching of grades is termed *gaming, manipulation, judgment,* or chance, the conclusion is the same: passing rates were biased upward, overstated by the ways test-scores were measured relative to the pass-fail boundary.

Concerns about teachers' use of discretion in grading, especially when grading their own students' exams, and the obvious conflict of interests, have combined with concerns about outright cheating to spur remedial measures. Quite apart from illegal cheating, however, four New York City high schools—in order to achieve objective, dispassionate, grading—had been compelled to have their scoring "supervised" by someone from another school.[46] How effective hat supervision has been remains unclear, as does its added cost.

Legal Gaming of Test Scores: Who Take the Tests?

Legal gaming is the most important but least understood and recognized way for succeeding by the test scores. Another is not distinguishing among high school graduates in the State of New York who do and do not deserve the Regents Degree award. Actions that are both legal and consistent with the educational or other social goals of federal legislation are surprisingly common. They are also the most

[43] Martinez and McGinty, "Students' Regents Test Scores Bulge"
[44] Martinez and McGinty, "Students' Regents Test Scores Bulge"
[45] Martinez and McGinty, "Students' Regents Test Scores Bluge"
[46] Martinez and McGinty, "Students' Regents Test Scores Bulge"

problematic forms of gaming simply because, being lawful, they may be overlooked by State regulators, the press, and the electorate. They "fly under the radar" -- difficult to uncover and eliminate.

Gaming, lawful or not, though, is by no means overlooked by stakeholders -- school administrators, teachers, parents of students, state government officials, and others pursuing their self-interests, all have stakes in the process of designing and grading exams and, more generally, rewarding "better" performance. Stakeholders' responses to incentives often conflict with social goals. The more powerful incentives are for schools and teachers to improve their students' measured test-score performances, to qualify for the rewards. Because legal gaming avoids the penalties for cheating or other forms of illegal gaming, and because of the incentive to nudge test scores upward, to justify treating a score as passing if it is "close" to qualifying a student for a significantly greater reward, legal gaming is more appealing and likely more common.

Legal Gaming is Hiding in Plain Sight

Legal gaming of schools causes test-score statistics to be more powerful they are intended to be. With gaming, a school, school district, or another educational unit seems to be performing well even when it is not. High stakes testing, for example, may appear to be obeying laws and regulations, yet undermine one or more of the educational accountability goals.

The key to "successful" legal gaming in a school is recognition of a simple fact: test-scores depend on *who does and does not take a specific test and who grades it.* It turns out that there are many ways to affect the actual test-taking pool, and they lead to more of the higher-performing students taking the tests, and more low-performing students avoiding the test-taking; the results are the raising or lowering of average class scores and passing rates.

To make clear both the variety of legal gaming mechanisms and the involvement with them of numerous stakeholders, including students, their parents, teachers, and school administrators, I turn now to this section on some strategies that are not generally considered gaming but that are, including: (a) "*red-shirting*" – parents delaying a student's start of elementary schooling, so as to avoid being one of the youngest students in the class and for future school years; (b) parents *opting-out* their children from taking "compulsory" standardized exams and, if the student is averse to such stressful testing; (c) parents *home-schooling* their low-performing children, and (d) disproportionate expulsions from school of low-performing students who would otherwise be likely to depress K-12 schools' average class-specific grades. All these devices are legal; all can be used to exaggerate a school's accomplishments.

Legal Gaming by Parents

Parents confront a number of choices for their school-age children, options differing in their expected rewards and penalties. Parents can influence the age at which a child starts school, whether the child goes to a conventional school, is home-schooled, or is simply a dropout, whether a child who is enrolled in a public school takes the standardized tests or avoids them through parental "opting out" of a child from normally mandatory exams, and whether a child remains in school beyond the minimum school-leaving age. Each decision has implications for class proficiency measures, and for federal need-based grants to a school.

As rewards and penalties change, so do stakeholder incentives and choices. The greater the rewards for better performance on standardized tests, for example, the stronger are parents' incentives to delay the age at which their children start school, so they will likely do better on the tests, and the

greater the appeal of home-schooling is for low-achieving youngsters. Rather than exposing children to tougher competitive pressures in formal schools, homeschooling is attractive to high-achieving students whose parents object to the tightening curricular standards of public schools.

Parents clearly influence their children's schooling actions, although they are by no means the only stakeholders facing incentives to choose one or another educational path. Moreover, a particular stakeholder group may well be subject to conflicting incentives tugging in opposing directions. A school principal, for example, can be under simultaneous pressures to encourage parents to assure that their high-performing children take the standardized tests even if the low-performers do not.

At the Brooklyn New School, a public elementary school in New York, pressures from students and parents complaining of excessive standardized testing led another stakeholder, the principal, Anna Allenbrook, to make it easy for students to engage in alternative activities on a testing day, thereby avoiding taking the exam scheduled for that day. More than 550 New York State school principals had signed a letter to parents expressing concerns about standardized tests, their financial strains on schools, and negative student reactions such as crying and vomiting.[47]

By contrast to these responses to pressures, principals in Portland, Oregon public schools (PPS) responded to another set of stakeholders, school administrators—to discourage parents from withdrawing, their children from standardized testing. A PPS spokesperson, Christine Miles, emphasized that a greater number of opt-outs reduces the percentage of students taking the tests, allegedly causing one elementary school, Vernon, to be ranked two spots lower in the state than it would have if there had been fewer opt-outs of higher-performing students.

Attributing gaming opportunities to any single stakeholder group is clearly a simplification. But it is a useful one for understanding the basic processes through which school gaming occurs and is rewarded. Here are eight gaming instruments that are largely under parental control:

(1) "Redshirting" (Delaying) When A Child Starts School Or Sports Training. Consider a parental decision on whether to have a child early or late in the year, which will later affect the child's school-starting age. Parents of a child born late in a calendar year—say, in November or December, may have the option of beginning school and sports training that fall or delaying it for the following September; choosing the delay would make the child one of the oldest and probably tallest and strongest in the class, while choosing the early school start, perhaps beginning when the child was a few months short of his or her fifth birthday, would likely make "him" one of the youngest. Postponement would make the child not only one of the oldest in the class, but physically stronger, taller, emotionally more mature, and, in the context of a potential career in sports, say in ice hockey, more advanced than his classmates.

Such a delay is analogous to the collegiate football practice known as "redshirting" -- holding a freshman player off the team for a year, but allowing him to learn the playbook and participate intramurally, wearing a red jersey sweatshirt during practice scrimmages—so that his fourth (and final) year of eligibility can extend to a fifth year of college, when he will presumably be a better athlete.

To the extent that parents gauge their children's success in school, academic or athletic, by the ability to compete successfully with their classmates, and to perform better on standardized tests, parents of a child born late in a calendar year has an incentive to delay school starting. In fact, the

[47] Mead, "The Defiant Parents: Testing's Discontents"

redshirting of kindergartners by their parents has become increasingly popular. As recently as 1995, redshirted first and second graders constituted 9 percent of their classmates, but that leaped to 17 percent by 2008.[48] The result is that redshirted children are among the oldest in their classes, not just at the start of elementary school but throughout their education.

Perfectly legal, this is nonetheless a form of gaming—seeking to increase a student's competitive advantage. Research on the effects of relative age on a child's learning is not entirely consistent—some has found the academic advantage switches by college age—but parents nonetheless have that strategic choice.[49]

Parents are not the only driving force affecting young people's future opportunities. As writer Malcolm Gladwell pointed out for Canadian youth hockey players, when the calendar year of a child's birth determines eligibility for a particular level of competitive hockey, being born in the early weeks or months of a calendar year rather than late in the preceding year, is advantageous; a baby born in early January is essentially treated as a year "younger" when entering kindergarten than one born a few weeks earlier, late in December of the preceding year; and the older child tends to be a stronger prospect for hockey success—having an added year of potential experience, practice, and training.[50] Whether this advantage results from intentional parental gaming is not clear, but in Canada, where hockey is of considerable occupational importance, that cannot be ruled out.

(2) Opting Out of Standardized Exams. The NCLB and subsequent ESSA laws effectively required public school students in the U.S. to take annual standardized tests in reading and math, and teachers and schools were being held increasingly responsible for students' test performances. Such testing standards actually date back even further -- to the federally commissioned 1983 report on "A Nation at Risk," which portrayed a decline in U.S. public schools' test performance. Some states had adopted standardized testing, which brought opposition from parents who opposed government intrusion, and in the early 1990s some states began to allow parents to allow their children to opt out of statewide standardized tests for their children. When NCLB and its goal of *nationwide* testing became law in 2002, tension between the two principles—mandatory nationwide standardized testing but allowing parents to over-ride the requirement and allow their children to opt out of the testing-- accelerated. That year alone, some 50,000 to 75,000 parents, largely in California, chose to excuse their children from the high-stakes testing, and later ESSA explicitly authorized states to allow parents to opt-out their children from the exams.

California's standardized testing had begun earlier, in 1972, spreading through several programs over the decades. Opt-out rules were enacted by the state in 1995. In 2001 at the San Diego high school, Scripps Ranch, 118 students (7.9 percent) were opted-out of the state's standardized tests by their parents—in the aftermath of a row over the principle of allowing those waivers, and the school's overall average test score dropped. Those students who were opted out were believed to have been not a random sample of students but disproportionately high achievers, and their nonappearance on exam days brought down the schools' test-score averages, harming the school's status as a "distinguished" or "blue-ribbon" school that attracted new residents to the area. The inverse relationship between the

[48] Konnikova, "Youngest Kid, Smartest Kid?"
[49] Konnikova, "Youngest Kid, Smartest Kid?"
[50] Gladwell, *Outliers: The Story of Success,* 1: 15–34.

number of a school's test opt-outs and the subsequent declines in average test-scores had been documented for a number of California schools.

While parents had made the formal opt-out requests, teachers had become increasingly active in supporting and encouraging parental actions, reflecting growing opposition to the P4P movement; even in the 1990s, student test-scores affected their teachers' pay.[51]

Test opt-outs and their links to standardized testing and teacher rewards have not been limited to California. New York State, one of the earliest states to adopt tests aligned with the more difficult Common Core (CC) standards, witnessed a dramatic increase in opt-outs within one year after the CC testing standards were introduced in 2013. There were ten thousand student waivers from the CC test that year, but when the tougher standards and increased student failure rates became clear, parental backlash sharply increased and student test waivers the next year vaulted to sixty thousand.[52] This jump came in the aftermath of student complaints that the new tests were too long, and parental complaints that the tests were making it harder for their children to graduate.[53] By 2021, a growing number of parents nationwide were not only opting out of state testing of their children but also "refusing to allow their children to take standardized tests."[54]

These events lead to two clear conclusions:

1. Higher test scores are not equivalent to increased learning.
2. What we know is how to measure test scores, not "learning."

(3) Students Missing Exams: Opting Out. In public schools, where testing had been required, the consequences of a student missing an exam, claiming "illness" or something else, varied. But unless opting out is permitted under state law, a student missing a standardized exam must be permitted—some would say, required—to take a "makeup" exam and be included in class and school performance statistics. Otherwise, those statistics could be gamed by facilitating, if not even mandating, low-performing students to evade testing.

If the same exam is given again, however, there are obvious test-taking security problems; and if a different exam is given, there are costs of developing new questions and assessing their comparability. Little is known about the numbers of students missing exams in states that do *not* permit opt-outs, but the incentive for students expecting to do poorly to miss exams is clear, as are their teachers' and principals' incentives to permit it, notwithstanding the uncertain effects on federal funding.

In any case the incentives to take or to miss a standardized test raised doubts about the consequences of "mandatory" standardized testing in public schools. The resulting performance statistics reflect competing incentives and gaming opportunities, with the result that test-score differences across schools, subjects, teachers, and even states may reveal more about the effectiveness of gaming than the effectiveness of the national P4P efforts.

Opt-outs are not disappearing, but they are fading away as an element of growing opposition to national education standards. In New York State public schools in 2015, some 20 percent of all students

[51] Golden, "Student's Dream, Principal's Dread"
[52] Bidwell, "Florida School District Opts Out"
[53] Hernández and Baker, "A Tough New Test"
[54] Chen, *Public School Review*

in grades three through eight "refused to take the state's standardized tests in reading and math," a number that did not even count students who skipped the tests for a "known valid reason, like illness." The 200,000 students who skipped the tests was four times the number in the preceding year, and was by far the highest opt-out rate in the country. School districts that had more than 5 percent of eligible students not taking the annual tests faced penalties in the form of lost federal aid, and the state could withhold funding.[55]

Divisiveness of attitudes toward the CC standards and the accompanying standardized national tests vary not only among states but also among school districts within a state. At the same time that the opt-out rate reached 20 percent in New York State in 2015, it was under 2 percent in New York City.[56]

(4) Home Schooling. Parents can do more than withdraw their children from standardized exams; they can withdraw them from public school entirely. Homeschooling is one viable, legal alternative for parents, and like private and parochial school options, it tends to undermine national educational standards, by permitting greater parental autonomy of choice among alternative contents of acceptable tests.

What is critical for public policy is to recognize that when these choices are exercised, the pursuit of *private*, in this case parental, goals, and *social* goals can conflict. Parents' decisions to home-school children or to transfer them to nonpublic schools including home-schools, affect not only the children but also the average test scores and proficiency rates for students remaining in public schools. In turn, those effects have repercussions for the drawing-power of more decentralized and diverse authorities as occurred when ESSA replaced the NCLB Act, transferring more control and flexibility to state governments.

These federal laws requiring standardized testing incentivized parents and school systems to favor more centralized standards and associated rewards in return for federal grants based substantially on students' performances on standardized tests. The subsequent decentralization of authority under ESSA weakened the mandate of standardized testing on specified subjects, as what amounted to compensation to states and their schools for remaining within an evolving consensus on the appropriate role of national standards for K-12 schools and for their financial rewards for improved student attainment of the standards.

Parental choice among lawful alternatives is not new, of course; but mandatory standardized testing and compensation for better performance have attracted increasing attention of political leaders, students' parents, and other stakeholders in strengthening the role of education in out social system. What is new is that as the test contents for measuring "learning," developing and rewarding its curricular underpinnings, and developing standards for advancing the breadth and depth of education in an advanced economic system have become more complex and ever-changing targets. Establishment of educational goals that reflect clarity, pliability, measurements that are inexpensive to quantify but costly to game are hurdles slowing if not diverting our progress. The continuing struggle to define a meaningful set of "Common Core" (CC) educational standards in a heterogeneous and changing society is a multiplicity of hurdles where parents have many opportunities and incentives to withdraw ("opt out")

[55] New York Times Editorial Board, "Opting Out of Tests"
[56] New York Times Editorial Board, "Opting Out of Tests"

school-age children from one public school and transfer to another, or to pursue other routes for bypassing the current CC standards.

Parents are doing just that. In a process not ordinarily thought of as gaming the standardized testing regime, but that brings the same results, home schooling is becoming increasingly popular, albeit from a relatively small base. Federal government statistics show that the number of home-schooled children doubled from 850,000 in 1999 to 1,690,000 in 2016, and rose even more sharply during the COVID-19 pandemic of 2020-2022. Homeschooling in the U.S. is largely unregulated, which offers parents expanded freedom to choose educational content and to flee nationally standardized testing. Eleven states do not require families even to inform a school district or other state agency that they are home schooling, fourteen states do not mandate any subjects to be taught at home, and only nine states require parents doing home schooling to have even a high school diploma or its equivalent. In half of all states, home-schooled children are not required to take any standardized test or be assessed by any external evaluator.[57]

As governmental oversight of home schooling has relaxed, incentives have changed. In October 2014, Pennsylvania, a state that had previously required home-schooling parents to register annually with their local school district, to present a study plan, and at the end of the year to submit a portfolio of the student's work to a private evaluator, sharply cut its oversight. Out went the child's portfolio requirement; in came parental authority to certify that their child met high school graduation requirements and to "issue home-grown diplomas without any outside endorsement."[58]

The new Pennsylvania law retained some oversight, requiring a private evaluator—typically a former teacher who is paid by the family—to review students' portfolios. But as one critic observed, "it only takes one person who meets the qualifications who is willing to sign off on anything."[59] The potential for gaming this "oversight" relationship is not difficult to see, given the interest of the family in continuing the home schooling, and the interest of the supposedly independent evaluator, who, being paid by the family, has the incentive to develop the reputation for satisfying parental wants. The evaluator's conflict of interest is, in effect, that of a "double agent," for the family and for the state's home-schooling regulator, which does not generate confidence that the government regulator will dominate.

Other states followed similar deregulation paths. Between 2012 and 2014, Iowa, New Hampshire, and Minnesota cut their requirements that parents providing home schooling register with local school districts, and Utah eliminated academic subject requirements for home schoolers, ending the requirement that families file annual affidavits stating the intention to homeschool their children.[60]

Individual schools and their teachers and principals, not just parents, can game the test-taking process to demonstrate their performance. Many avenues are available to influence the "quality" of their student bodies and thereby their expected test-performances, to determine the allocation of class time to education rather than to test-taking skills, and to influence the number of students passing standardized tests.

[57] Rich, "Home Schooling: More Pupils, Less Regulation"
[58] Rich, "Home Schooling: More Pupils, Less Regulation"
[59] Rich, "Home Schooling: More Pupils, Less Regulation"
[60] Rich, "Home Schooling: More Pupils, Less Regulation"

(5) Test Preparation: Teaching to the Test. One form of allocating (manipulating) use of class time in ways that affect student performance on test scores is largely, though not entirely, under the control of teachers. The ties between student test scores and teachers' rewards produced a peak of contentiousness in New York State in 2010. State education officials, concluding that the standardized tests in reading and math had become easier to pass during the previous four years—indeed, *too* easy—announced that the exam scoring would be recalibrated in a way that would be "almost certain to mean thousands more students will fail." [61]

The goal was to lift student achievement by making it more difficult to pass the standardized tests, presumably challenging students and teachers to work harder and more effectively. There are many ways to make passing an exam harder, one of which is to make test questions truly more challenging; but that would be costly to develop. Recalibrating the scoring, rather than rewriting the test, is another, and simpler, way to make a test harder to pass. Either way, an increased failure rate would undoubtedly create pushback—from parents who saw more of their children failing key exams and not graduating on time, and from teachers who saw their own performance being questioned. Thus, the stated goal of the state education leadership, to raise standards, was unlikely to be aligned with the goals of teachers and parents, or even of school principals, who do not view higher student failure rates, even if "temporary," as success.

Four years after New York State trumpeted its determination to "recalibrate" and toughen tests and educational standards statewide, New York City teachers in a typical middle school were still spending part of nearly every day focused not on fostering student creativity and understanding, but on test preparation. In reading, this involved working with "sample passages, questions and review sheets mirroring the actual tests," and in math, students "did work sheets full of word problems from previous tests," and as the critical three days of state math exams approached, "students concentrated each day on test content, including double periods twice a week for what felt like 'every second' of class."[62]

Teaching-to-the-test is a tried-and-true way for teachers to increase test-passing rates—for example, by telling students the answers to questions on recent tests, or how to answer questions of the types likely to be tested. Whether the difficulty of test questions, determined at the state education department level, is increasing faster than teachers' adjustments of their test-preparation methods remains to be determined. The direction of those adjustments, however, is not in doubt: as long as teachers and principals continue to be rewarded for having more of their students pass the standardized tests, teaching to the test will be a powerful form of strategic gaming by teachers and perhaps principals—an unintended but predictable effect of the increased test difficulty. Whether the New York State tests were actually made more "difficult" is not directly observable, but the incentives for teachers and schools surely continue to encourage test-prep efforts that boost passing scores.

Citywide, teachers commonly report that "a decade of accountability-based reforms have deeply changed the classroom." Despite the New York City school chancellor's denunciation of the overemphasis on test preparation, there is no evidence of its reduction. Even after a new state law limited test-prep time to 2 percent of the entire school year, students at one intermediate school estimated that they devoted more than two-thirds of their English classes to preparing for the reading test, although there are ambiguities: Was the 2 percent limit for the entire year met by limiting the

[61] Medina, "State's Exams Became Easier to Pass"
[62] Baker, "Test Prep Endures in New York Schools"

heavy reading-test preparation to just a few days? In addition, when is class time test-prep, and when is it learning English? Since state law bases teachers' evaluations partly on their students' test scores, and those scores also affect students' admissions to competitive middle and high schools,[63] teachers, students, and parents share the incentive for higher test scores, and so share the incentive to meet the 2 percent regulatory limit on "test preparation" by choosing to interpret as time devoted to "studying English" what others, including state education officials, might regard as test-prep.

(6) "All" Students Must Take the Standardized Tests – Really? There is another regulatory constraint on the testing process that seemingly invites gaming by teachers and principals, and quite possibly with the involvement of parents. It refers to who *must* take the tests. Although the NCLB purportedly applied to *all* public school students, it actually mandated that "not less than 95 percent … of students" take the reading and math tests for each specific grade.[64] This 95 percent rule gave schools two avenues for raising average test scores and passing rates by manipulating the composition of the test takers: influencing who is counted as a "student" in a given class, and which of them are permitted, or even encouraged, to be among the maximum of 5 percent not taking a standardized test. The rewards for teachers, principals, schools, and, often, students and parents depend on average scores and "proficiency" rates, which are determined by the scores of those who take the test, and those scores are affected by whether more high- or low-performing students are included in a class total, and in the 5 percent or fewer who are excused from taking the exams.

(7) Excluding "Special Education" Students. Students with significant learning disabilities deemed by school officials to qualify them for special education services were typically excluded from regular class testing under the federal legislation (although they might be tested in other ways), and their exclusion would clearly raise the measured average class performance on standardized exams. In earlier times, a school's decision (quite likely with parental input) on whether or not to deem a student eligible for special education services depended on two kinds of judgments—the student's learning abilities and the added costs to the school district of developing and implementing an Individual Educational Plan (IEP) for the student. But NCLB changed the calculations.

Under NCLB a school's assessment of the benefits and costs of special education status for a student added something new—the effects of the decision on class and school test-score averages, a common measure of performance that influenced teacher and school administrator rewards. Because a special education student could be excluded from the regular class testing and performance statistics, the class average statistics would rise or fall depending on whether a student was or was not deemed eligible for special education services and test exclusions. The magnitudes of such marginal reclassifications by schools are not known, but the advent of federal legislation was a new force affecting public schools' and parents' incentives to, in effect, opt-out more children from standardized testing by nudging them *into* special education programs and thereby *out* of the testing pool.

Holding equal other forces affecting test scores, the incentives established by federal legislation would have stimulated increased numbers of special education students. For two reasons: schools could elevate their students' test score averages and proficiency rates, and parents of children with moderate learning disabilities could protect their children from criticism that they were depressing school performance statistics. With both schools and parents sharing this interest, growth in special education

[63] Baker, "Test Prep Endures in New York Schools"
[64] NCLB, §1111(b)(2)(I)(ii).

program participation would amount to a steppingstone toward better performance on the standardized tests.

But the actual pattern of change was not so clear. Even before the inauguration of NCLB in 2002, provision of "special ed" services was already growing. In 1993–94, 12.1.percent of children aged 3–21 were served by programs under the *Individuals with Disabilities Education* Act (IDEA), and by the school year 2001–2 that had increased to 13.5 percent, having crept up each intervening year. The subsequent passage of federal legislation brought no apparent acceleration, as might have been expected, and the slow growth continued at the rate of about 0.1 percentage point annually. The Individual's with Disabilities Education Program peaked at 13.8 percent in 2004–5, subsequently declining slowly every year to a rate of 13.0 percent in 2011–12,[65] even though the incentives under NCLB for schools to raise their test score statistics were growing. The range of forces affecting special education program participation is not fully understood, yet the evidence does not suggest that schools and parents had been using special education programs as incentives to show improved performance.

(8) Expulsions and Dropouts. Expulsions are typically disciplinary actions by schools in response to student misbehavior, while dropout decisions reflect actions with joint involvement of students, parents, and schools. Under NCLB, dropping out of low-performing students had the same effects on class performance statistics as the expulsion of low-performing students. The incentive for schools to encourage low-performing students to leave the test-taking pools—to drop out—is the same as that to force those students out—to expel them. Schools have the incentive to use both mechanisms—to encourage low-performing students to drop out of high school and pursue a GED (General Education Development) certification, and to expel low-performing students for longer periods when they have been engaged in misbehavior; in both cases the result is improved class test score averages, other school performance statistics, and the accompanying rewards.

Encouraging low performers to drop out of school is more than a possibility. In 2004 the director of the World of Opportunity Adult Education program in Birmingham, Alabama, Steven Orel, reported that standardized exams encouraged schools to "write off" the most difficult students. He later noted that following NCLB, sixteen-year-olds started showing up at his GED program, then run by the Birmingham Public Schools, carrying documents saying they had "withdrawn" from the local Woodlawn High School. The forms, signed by Woodlawn officials, reported the cause as "lack of interest." But "kids were coming to us within a week or a month of leaving high school," said Orel. "It defied logic to me, he said: Why were these kids coming to me if they lacked interest?" [66]

A Birmingham school board member, Virginia Volker, provided an answer, after learning that 5.6 percent of the high schoolers, some 522, had "withdrawn" from school during a single year. She found the students had been "told to leave school after Feb. 15, when the state calculated reimbursement levels based on enrollment, but before April, when they would have taken the Stanford Achievement Test and could have dragged down their school's scores."[67]

The incentives were clear. By astute gaming of the dropout *timing* of low-performing students a school could obtain the annual state grants for students *before* they dropped out, and also increase class-wide test score averages *after* those dropouts. Financial incentives encouraged schools to retain

[65] Calculations from U.S. Department of Education, *Annual Report to Congress*.
[66] Schemo, Ninth Grade Key to Success"
[67] Schemo, Ninth Grade Key to Success"

their weak students long enough to gain the state's support, and to use their influence on parents of high-performing students to *discourage* their dropping-out.

Expulsion of a student is equivalent to the student's temporary dropout in its effect on test score statistics; in both cases the student does not take the standardized tests, and so, if a low performer is involved, test score averages increase. The key difference is that while dropping out is the decision of a student and parents, expulsion is an administrative decision largely controlled by the school, which can exercise its disciplinary control to influence test score performance —that is, passing -- rates. Expelling a student who is expected to do poorly on forthcoming standardized tests is not equivalent to short-period expelling of a high performer.

The reward system, by encouraging higher test scores, drives a wedge between the two cases, implicitly encouraging tougher expulsion penalties to lower-performing students. This incentive to gaming was not intended by federal legislation to reward schools' measured performance, nor to emulate the strategic penalty structure that had been adopted in Florida in the 1990s; it did, though, generate a predictable pattern of systematic expulsion incentives.

Research has disclosed strong evidence that Florida public schools responded systematically and predictably to the State's *mandatory* standardized tests. Evidence was found of selective, strategic, expulsions of students involved in inappropriate behavior. The schools, rewarded for students' higher test scores, used differential disciplines for student misbehavior in ways that bolstered test score performance statistics. When two students were expelled for a single event such as a fight, the student who had performed poorly in prior standardized tests was expelled for a significantly longer period than was the high-performing member of the pair.[68]

Florida schools were studied for the four school years 1996–7 through 1999–2000, the first years after the state introduced its own high-stakes exams, the Florida Comprehensive Assessment Test, to evaluate a school's performance. The research goal was to determine whether schools use differential penalties for students of varied performance levels who were involved in the same misbehavior and who were suspended. The central question was whether the timing and duration of suspensions differed systematically between the two students in each pair, and if so, whether their penalties differed in ways that had the most favorable effect on the expected average class test scores. In short, were the lower-performing and higher-performing students in each pair suspended differentially and in ways that had the most favorable effects on class test score averages? Was there evidence, in short, of gaming?

There was. Suspensions were examined in 41,803 incidents in which two students were suspended and the prior test scores for both students were known. Statistical controls were added for the fixed-effects of the incident involved, and for a number of other variables that might have confounded the interpretation of differential expulsion penalties—including race, gender, whether the student had been expelled previously, and was poor enough to be eligible for free lunch.

The finding: schools were consistently giving *longer* suspensions to low-performing students for similar infractions. Further support for the view that expulsion policy was being employed strategically came from another finding: the gap between the durations of expulsions for the two students in each pair grew substantially during the four years, suggesting that schools were either responding to

[68] Figlio, Testing, Crime and Punishment"

increasing rewards for higher test scores, or were learning how to be more effective in gaming test score statistics and their accompanying financial rewards.[69]

There was another indicator of the gaming of penalties according to their effects on the testing system and its test-based consequences. Differential expulsion durations would have effects on test score statistics only in years in which the tests were given—in grades four, five, eight, and ten—and not in other grades. Thus, there were incentives for differential expulsion durations for students in some years and not in others. If expelled students were in a grade not tested there was no incentive to expel them differentially, to raise test score averages.

The expected results were again found; differential disciplinary suspensions were observed only for students in the testing grades—where they mattered. School incentives were clear—to have as many high-performing students as possible in school when the tests were to be given, and to have low-performing students stay away at those test times, to raise class test score averages. "Students receiving long suspensions during the testing window were more likely to miss the examination and its make-up dates." And longer suspensions were given to lower-performing students, thereby again raising test score averages by allowing high performers to more-quickly return to the set of students taking the exams and generating greater rewards. The result of the systematic use of varied expulsion durations was to favor high performers with shorter penalties and, and thereby to raise test score averages and rewards**.**[70]

However understandable the schools' expulsion strategy was, it intentionally distorted the test-taking pool and raised the resulting performance statistics. Using such selective expulsions-- gaming—to improve measured performance was surely not a method intended by the framers of the Florida school-accountability system.

Legal Gaming by State Governments: The Common Core Standards and Strategic Relationships Between States and the Federal Government

During the more than a decade of NCLB, each state could qualify for federal support that depended substantially on its students' rates of "passing" that state's standardized tests in math and English, and on its annual "progress" in increasing those passing rates. Each state had been free to decide on the specific content of its tests and on what constituted a passing score. Starting in 2014, this began to change, in ways with potentially enormous consequences.

The new element was the ascendance of what amounted to *national* standards -- a "Common Core" (CC) of curricular goals, their alignment with standardized tests, and Federal financial awards for "good performance" on those tests. The decentralized authority of each state to decide on its tests' contents, on what levels of student performance constituted "proficiency" in each subject, and on how to deal with the power of the U.S. Department of Education to tie its financial aid for a state's K-12 schooling to the state's acceptance of the Common Core standards of knowledge.

That tension between the Federal government's power of the purse, and the diverse state governments' views about what should constitute appropriate educational content, continues, but has begun to dissolve. Now, each state that agrees to adopt the CC *national* standards, which continue to evolve, is linked to both Federal grant support and to adherence to the *national* CC content standards.

[69] Figlio, Testing, Crime and Punishment"
[70] Figlio, Testing, Crime and Punishment"

Authority over test content and standards for meeting or exceeding them have been shifting away from a uniform set of national standards and rewards, in favor of decentralized state and local authorities. With changing consequences and gaming opportunities.

A power struggle between the federal and state governments continues, in addition to the tussles between state and their local schools, triggering changes in incentives throughout the diverse school districts. As of April 2020, of the 45 states that at some point had adopted the national Common Core standards, 24 subsequently repealed or revised their approvals; and four states have entirely withdrawn – Arizona, Oklahoma, Indiana, and South Carolina – while Alaska, Nebraska, Texas, Virginia, and Puerto Rico never adopted the standards, which are still evolving.

The Common Core concept and its implementation may seem to have little to do with the perils of P4P in public K-12 schools, but that conclusion would be wrong. Gaming of a social system reflects, at its base, opportunities for participants in the process and its financing to take actions that, even when legal, undermine the program goals. We have already seen how parents who opt-out their children from the standardized testing, school administrators who expel troublesome students more severely when they have been performing poorly on standardized tests, and high school teachers who grade tests more leniently when that is necessary for a student to attain the threshold for a "passing" grade. All of these results were legal, all were *un*intended under NCLB, and all were counterproductive as they undermined social goals.

There is yet another participant-stakeholder in the education process, its testing and rewarding, that can game the system: the individual states, territories, and District of Columbia. The U.S. Constitution, which does not give explicit responsibility for public education to each state, does provide opportunities for a state to engage in what amounts to a bilateral bargaining process in which a state's acceptance of federal funding for "education" is negotiable. So not only can parents, teachers, and administrators game the education system reward structure, so can the states. And they do—often legally.

The clash between the Common Core goal of developing *national* education standards and rewarding their achievement as measured by standardized tests, ensures tension. It also ensures efforts by each state to search for ways to reap larger federal grants for aligning its standardized exams with the national CC standards, while retaining the state's independence and diversity – another recipe for gaming.

The contentious debate over CC standards—which were developed in response to a joint initiative of two interstate (but not federal) organizations—is the latest forum for the P4P debate in K-12 schooling. The concept of a Common Core of material to be learned by all students and rewarded according to their performance on standardized tests, is inconsistent with each state's control over its educational curriculum and its standards for demonstrating proficiency. The struggle is still evolving as states decide whether to support the latest version of the Common Core standards, or take an independent path with fewer restrictions but fewer financial rewards from *the* Federal Government. Even states that have made the choice, whether agreeing to adopt the CC standards or not, can and do change their positions as those standards evolve and their consequences became clearer.

This process, still in flux as of 2022, makes clear the challenge that is fundamental to a P4P regime in any industry, not merely K-12 schooling: defining a program's s*uccess* -- translating it into a measurable form, rewarding it, and avoiding counterproductive gaming. These are not straightforward,

one-time, decisions. States and territories want larger federal grants, but not the accompanying constraints on their schools' curricula; the federal government, by contrast, wants to establish national education standards but must win over, or "bribe," recalcitrant states to accept the federal standards and their enforcement. Government is the chief barrier, for each state may opt-out of the CC compact or enter and re-enter it, so state governments are the primary hurdle in the federal restriction on states' access to federal grants. The interactive skirmish became more evident as states entered and exited from the CC compact.

Early responses to the concept of establishing CC standards of educational success were strongly supportive. In 2008 the Bill and Melinda Gates Foundation announced its allocation of $233 million to develop and promote advancement of national educational standards, which were being developed in a bipartisan effort by the National Governors Association and the Council of Chief State School Officers. Forty-five of the 50 states, and the District of Columbia, soon supported the CC standards and their associated standardized tests. For math, reading, and writing, goals were established requiring fifth grade students, for example, to be able to write essays in which they demonstrate their ability to "introduce, support and defend arguments, using specific details."[71] But states and territories have joined and quit the shifting compact.

By 2021 the CC standards had been adopted by 42 states, the District of Columbia, the Department of Defense Education Activity (a civilian agency overseeing the education of military offspring both at home and abroad), and the territories of Guam, the Northern Mariana Islands, the American Samoan Islands, and the U.S. Virgin Islands.[72] As the commitments associated with the Common Core evolved, states and territories have joined and left the CC membership; but as I noted earlier, state holdouts remain.

The initial massive expressions of support for a national CC curriculum soon began to wither as the restrictive implications of uniform national K-12 education standards became clearer. The first defection came in August 2014 when Indiana, one of the first states to have signed onto the "final" 2010 draft of the CC standards, became the first state to rescind its support. There was no single explanation for that action, but the desire for local control was a major force.[73] Within months, Oklahoma and South Carolina also defected, and at least 15 other supporters have been wavering in their support: Colorado, Idaho, Iowa, Maine, Missouri, Montana, Nevada, New Jersey, New York, North Carolina, Ohio, Oregon, Pennsylvania, Utah, and Wisconsin, and in Massachusetts grassroot opposition has developed among groups of parents, educators, and local elected officials. So, all of the five states originally *opposing* the CC and its tests and rewards have sustained their opposition, three of the original *supportive* states have withdrawn their support, and some 15–16 other initial supporters are reconsidering as of 2021. Divisiveness, not consensus, appears to be growing,

The Department of Education and other CC supporters will doubtless continue to try to smooth over the states' diversity that has generated much of the opposition to national schooling standards, in order to expand acceptance of *some* set of CC standards and administrative interpretations. But their effectiveness is questionable. Despite the attraction of financial incentives, the heterogeneity of states'

[71] New York Times Editorial Board, "Caution and the Common Core"
[72] Common Core State Standards Initiative, "Standards in Your State"
[73] Banchero, "Indiana Drops Common Core"

social, cultural, and political views can be expected to repel adherence to uniform national educational standards and their testing and rewarding.[74]

To see the fundamental forces working to erode support for national education standards, imagine conditions if all states were to agree to the current CC standards and their associated tests, rewards and penalties. Surely, individual states' test performances would be compiled and made public at some point. Some state would be first in the test-passing race; another would be last. States likely to do relatively poorly have the incentive to avoid the adverse publicity and comparisons by opting out of the CC standards altogether -- or, at least, by pushing for waivers that delay, perhaps for years, implementation of penalties. Some states may retain the freedom to adopt their own educational standards, tests, and standards of passing, albeit at the cost of losing some federal support. A state's opting out of CC testing standards is a form of gaming analogous to a parent's test opt-out decisions for a child; both bring results that parents prefer but that undermine the education program's broad and changing goals.

Summing-up: Who Games, and With What Unintended Results?

The wide scope of gaming and the legality of many of its forms, all of which impair national educational goals, highlight the difficulty of avoiding gaming in K-12 schooling. Incentives to game are complex, since every participant in the education process and its financing has an incentive to engage in some form of gaming—illegal, legal, or borderline—under particular identifiable conditions.

In the case of the Atlanta public schools, it was teachers and low-level administrators who were illicitly correcting students' incorrect test answers. In the case of influencing low-performing students to drop out, as a mechanism for raising class-average performance scores, school administrators, teachers, and parents took the lead. In the case of opting-out of standardized testing, it was parents—but encouraged by teachers and permissive state laws. In the case of strategic suspensions of lower-performing students it was school administrators who were meting out the timing and durations of mandatory absences from schools.

A comprehensive perspective on gaming must recognize not only its many forms, but who the stakeholders are, which of them engage in gaming, and what their incentives are. So must an effective strategy for limiting or even eliminating gaming, which varies with the industry, its technology, and its regulation, as will become more apparent in our journey through industries that are concentrated in the public, government, sector, such as in public schools and courts, or the private *nonprofit* sector, such as philanthropy, and private firms pursuing their own profit but subject to a variety of regulatory constraints that vary with the enforcement budget.

Every stakeholder in the schooling process has an incentive to affect the outcomes of the gaming process: a parent wanting to advance a daughter's education will want to maximize her scores on standardized tests to strengthen her college application; a K-12 school system will want to maximize its budget from the state, school district, and federal government; and a particular teacher will want to see her students' test-proficiency rating increase if that will favorably impact her long-term earnings. These incentives differ, and may even conflict.

[74] Weisbrod (1975) examines, at a theoretic level, relationships between diversity and demand for any public good, and the responses of the public and private nonprofit sectors.

The cheating on test scores in Atlanta, for example, definitely improved performance as *reported,* for students, their teachers, and school administrators. But unintended side-effects followed as a result of the systematic over-reporting.

At least one school's students were allegedly performing so well that the school lost governmental financial support for its "low-performing" students.[75] As performance was measured and reported, there were virtually *n*o low-performers, but reality was different. There were, in fact, low-performing students, and they did suffer as a result of their schools' gaming -- under-reporting the true numbers of low-performers, the result of over-reporting their test scores. So *reported* overall class performance rose, especially at the bottom of the distribution, but only because it had been intentionally overstated by teachers and administrators responsible for the grading and its standards.

The combinations of multiple forms of gaming and multiple participants in the test grading process yielded a vast potential set of gaming opportunities. Stakeholders in the education process could, and did, pursue their self-interests, but at the cost of decline in student learning.

Decisions by another stakeholder group, parents, may well be intended to help their own school children, but the effects can be otherwise. Choosing private schooling or homeschooling alternatives to public schools, for example, or, where feasible, keeping their children at public schools but opting them out of standardized exams—affect not only their children but also other students in their classes. Test-score class averages, and passing rates for the remaining students fall if, for whatever reason, high-performing students leave the testing pool, or rise if those leaving the testing pool are below-average performers.

Parents typically make the opt-out decisions for their children, but other stakeholders—teachers, school principals and advisers— also influence the choices. Teachers who oppose standardized testing because of its links to their salaries and promotions have the incentive to game the system toward their own ends, as well as to signal parental protest of high stake testing. Yet teachers may also have an opposing incentive—to *discourage* parents from opting-out their high-performing children, which would likely decrease class average test scores, increase class failure rates on standardized tests, and negatively affect teachers' compensation.

School Principals, however, may have incentives contrary to those of at least some teachers. So different strategic behavior can be expected, if principals' actions toward teachers reflect pressures from school boards wanting to demonstrate the excellence of their local public schools by providing evidence that fewer students are opting-out of standardized testing than had been the case under previous school leadership, or that high-achieving students were remaining in schools longer, boosting test score averages. So Principals' own rewards might conflict with other stakeholders being rewarded for *reduced* test opt-outs, if that were interpreted as a compliment to the school's drawing power, not as an instrument for influencing which students do and do not take the standardized tests in English, math, and science.

Principals also face pressures from other stakeholders, particularly from teachers and parents opposing high stake testing. And so, with principals' performance and rewards—salaries and

[75] Winerip 2013.

promotions—being determined partly by evidence of student, parental, and teacher satisfaction, and partly by test-statistics, it can be appealing to overstate test score performance.

Each stakeholder in the game of strategy has the incentive to advance his or her own interest while accounting for the likely responses by the other parties. Thus, while a teacher may have the incentive to encourage some parents to opt-out their low-performing children from the standardized testing, the teacher typically has no incentive to inform parents of the conflict of interest between her role as an advisor of parents and students, and her role as agent for advancing her self-interest. The role of a double agent, in a school or elsewhere, is never easy.

Principals also face conflicting incentives: to encourage teachers to show a school's success by minimizing opt-outs and dropouts, while simultaneously achieving excellent, even improved, class performance on standardized tests. The two goals conflict: higher average test scores can also be achieved by the exit of low-performing students from the testing pool.

A typical state also wants both low student opt-out rates and high average test scores—plus a third measure of achievement, obtaining larger federal grants for its public schools. A state thus has the incentive to sign onto Common Core educational standards, to qualify for at least some federal funding, but then to fight for weakening those standards or receiving postponements (waivers) of the state's implementation of the standards, knowing that it can always change its decision and opt out of the standards, as numerous states have done.

What Is "New"? Reactions to High-Stakes Standardized Testing, Past and Present

Given that all these forms of test-score gaming—by parents, teachers, principals, school boards, and state and Federal governments—are products of high-stakes P4P incentives—where the stakes differ among stakeholder groups, there is at least one safe prediction: increased rewards for better measured performance in K-12 schools will bring increased resistance from some stakeholder group. They are already. And we can also learn from the past to better view what might be in store for the future.

In 2014, twelve years after NCLB became law, there were 179 bills in state legislatures to curb standardized testing and its rewards and penalties. Five national groups instituted "Testing Resistance and Reform Spring 2014" as a guide to help parents organize local protests, opt their relatively low performing children out of these exams, and restrict states' support of high stakes testing. Politically conservative groups joined with teachers and parents in opposing standardized tests.[76] It was in reaction to that movement that congress replaced the NCLB with the more flexible ESSA in 2015. Yet the basic reliance on standardized test scores continues as a way to gauge how well a school and its teachers and principals are performing and whether they are "improving," with rewards and penalties accordingly.

The history of P4P in schools is a long one that has had ups and downs, but the clash continues between the appeal of *stronger* incentives, to improve education by offering rewards to producers for more and better productivity, and the appeal of *weaker* incentives, to discourage development of reward systems that discourage actions that "succeed" in the sense that producers of "more" and "better quality" products reap greater rewards without really succeeding in increasing performance. Is the struggle different than in the past?

[76] Banchero, "Indiana Drops Common Core"; Banchero, "Standardized Testing."

In the education realm, previous efforts to measure and reward performance have typically ended in disappointment if not outright failure. Beginning in 1969 the focus of P4P in government was on "performance contracting." The Federal government acted on the premise that private firms, lured by profits based on test scores and job placements, would be more efficient than government in improving students' performance in education and manpower training. All that was needed, it seemed, was to emulate the private market model—to provide stronger incentives for firms for their measured effectiveness. With federal government support,

> … over 150 school districts contracted with companies to deliver instruction, as the Nixon Administration initiated a vast privatization experiment in Texas and Arkansas. None of these performance contracting experiments significantly improved instruction. After charges of corruption, teaching to the test, and a lack of productivity results, the program was abandoned and the single salary schedule continued to dominate school districts throughout the U.S.[77]

The federal government contracted with private for-profit firms to demonstrate that they could succeed where public schools could not. To provide the appropriate financial incentives, the government paid private contractors little or nothing unless their students' test scores improved by at least some government-specified amounts, such as 0.6 of a grade level in the course of the school year, and were paid more if test-score accomplishments were greater. This P4P accountability lasted some fifteen years, ending as a consensus developed that "obsession with test scores, stimulating excessive test preparation and 'teaching to the test,' had produced illusory progress and narrowing of curriculums to the most easily tested basic skills." [78]

President Nixon's 1970 message to Congress had proposed establishment of a National Institute of Education (NIE), to improve measurements and quantify "educational performance." The NIE, he said, was particularly to "pay as much heed to the 'immeasurable' aspects of schooling (largely because no one has yet learned to measure them), such as responsibility, wit, and humanity as it does to verbal and mathematical achievement."[79] It was clear even then that more complete measurement and valuation of school performance was needed if better performance was to be rewarded, but that this was bound to be elusive.

Now, with over 50 years of hindsight, that dire forecast remains well-founded. The ingenuity of school stakeholders in developing gaming mechanisms for *measuring* performance and *paying* rewards for their achievement have become substantial and widespread.

Measuring school performance is necessary for rewarding it, but what constitutes *good* performance? That is by no means clear, reflecting judgments about what the social goals of public education *should* be. In a 1988 address, Albert Shanker, then president of the American Federation of Teachers between 1974 and 1997, weighed-in on this critical foundation for any P4P or merit-pay system. He promoted schools as ways to advance "social mobility for working-class children and social cohesion among America's increasingly diverse populations."[80]

---

[77] WEAC, What Do We Know about Merit Pay?", 4
[78] Rothstein, et al, *Grading Education,* 28.
[79] Rothstein, et al, *Grading Education,* 29, (citing Nixon, "Special Message").
[80] Kahlenberg and Potter, "The Original Charter School Vision"

However, the goals of education were actually defined, measured, and rewarded, one conclusion was clear; they involved more than math and English. But Shanker did not tackle the issue of how, or even whether, to measure the expanding dimensions of performance expectations, let alone how these broader goals should be linked to an overall incentive structure. Neither did he address the distortionary consequences of schools, teachers, principals and parents being rewarded for successes in some dimensions of learning but not in others.

A Longer View History of Rewarding Performance in Schooling

Concerns about incentives and how to get them "right" in K-12 schooling are not new, as I showed above. The same hurdles we confront today—how to *measure* "performance," and how to *reward* it while *avoiding stakeholder gaming*—emerged in mid-nineteenth-century England. In 1858–61 a Royal Commission chaired by the Duke of Newcastle recommended that public support for education be tied to "payment by results." Robert Lowe, of the Education Office, who produced a code of education regulations and procedures in 1860, offered a Revised Code in 1862 to reflect the main findings of the Newcastle Commission report. Schools could henceforth obtain 4 shillings a year for each pupil with a *satisfactory attendance* record, and an additional 8 shillings if the pupil passed exams in reading, writing, and arithmetic.[81] But "after a 30 year try, British schools abandoned merit pay by 1900 due to cheating scandals, cramming, the growing influence of the testing bureaucracy, and the extent to which teacher concern about financial awards and punishment were warping the educational system."[82]

These forces remain with us today. In U.S. schools, measured performance continues to focus on math and reading, rewards continue to be tied to test scores largely on those subjects, but with the additions of **S**cience, **T**echnology, and **E**ngineering, in addition to **M**ath, to constitute today's "STEM"-oriented focus of college education for the labor market believed to lie ahead. Now, most other dimensions of education and learning remain largely ignored. The current focus on Common Core standards and their testing may or may not breach that frontier, but major subject-matter areas remain unmeasured and unrewarded, and the distortionary effects of gaming endure.

When President Nixon called for a new education agency to observe "immeasurable" dimensions of education along with the measurable ones, he did not say what the problem was --why, that is, was rewarding only the measurables problematic? Was it simply that the performance measures were incomplete, which they clearly were and still remain?

Or was there a deeper insight—that rewarding only a very partial set of measures of school performance was producing negative, though unmeasured, side effects from gaming? And was it recognized that those distortions, while likely small at the time, were becoming increasingly serious as rewards for measured test scores increased?

We cannot be certain what the thinking was in 1970, let alone in 1860. But the real problem, then and more so now, was not simply the incompleteness of measured performance, but the *distortionary*, counterproductive, incentive effects of *rewarding* some but not other dimensions of educational performance. If the social problem was simply the *incompleteness* of math and English as measures of *total* school performance, why would rewarding only the *measurable* elements of performance, but ignoring the *immeasurables,* be problematic, especially if the rewards were "small" Conceivably,

[81] Simkin, "Robert Lowe"
[82] WEAC, "What Do We Know about Merit Pay?", 3.

although it turns out, not likely, higher student scores on English, math, and science tests could be increasingly rewarded without cutting teachers' incentives to develop student capabilities in social studies or in their "wit" and "humanity," which remain unrewarded today?

It could, indeed, be argued that teachers' incentives to devote time and energy to these *un*rewarded, immeasurable, aspects of education could not be any weaker; rewards for them were, and remain, essentially zero. Incentives for teachers and schools could be disregarded, going no lower. Were that true, greater rewards for the measured math and English achievements under No Child Left Behind would have little or no effect on provision of "*unmeasurables."*

It is not quite that simple. Rewards and penalties that affect teacher incentives take many forms—not merely direct payments. Schools may be required, for example, by state regulatory mandates or school district rules, to devote a specified minimum amount of class time per week to a subject, such as art, even though there is no standardized measure of its performance effectiveness, and no associated explicit reward. As a result, more powerful rewards for math and English siphon resources away from the "unrewarded" subjects, especially in subtle ways that are costly to detect.

It is also important to recognize that rewarding any performance indicator is a two-dimensional process: the particular form of performance must be *both* measured *and* rewarded. Either one, alone, will not alter incentives. A school's or teacher's "success" in reducing childhood obesity, for example, is even more easily measured than the student's math prowess—although, to be sure, determining the value-added by the teacher is far from easily measured in both cases. But weight loss is not rewarded, and "teaching about it" and its causes are generally not mandated. So, schools and teachers have little incentive to devote scarce class time to "reducing obesity."

Efforts to reward teachers or administrators for anti-obesity "success" would encounter the same kinds of gaming incentives we have already shown to exist under NCLB: a teacher would have the incentive to attract more students who are just above the obesity threshold and then manage to reduce the student weight to just below that threshold; to garner the rewards, students whose weight was far above the threshold would be avoided, as would those already well below it. Typical of all "pass–fail" reward systems, a tiny difference in measured performance can lead to a large difference in rewards. The resulting incentive for teachers is so clearly inefficient that the easily measured "weight" or weight loss" has not become a form of *performance* that is rewarded. Doing so would have brought about the inefficient threshold-reward effect, or, to avoid it, complex judgments about appropriate incentives for rewarding each teacher and administrator for each student's weight loss and its retention.

Whether the unmeasured and unrewarded elements of success are in the arts, sciences, social studies, weight loss, or anything else, the more comprehensive the performance measures are, and the greater the rewards are to stakeholders, the more concentrated the effects would be of expanding rewards by shifting efforts away from the unrewarded activities. But a less sanguine result is that they would not.

In any case, expanding measurements and rewards beyond the math and English covered in the original federal legislation, later expanded to include testing of science knowledge and other subjects, raises the "opportunity cost" borne by teachers when they try to devote more time and effort to the remaining *unrewarded* subject. An *un*rewarded element of school performance might receive *some* attention if a teacher or Principal viewed it as sufficiently important to justify cutting time from the

rewarded subjects; however, the greater the rewards for measured performance, the larger would be the teacher's effective penalty for taking time from them.

The systemic differences between rewards for students' higher test scores in math and English, and later in science, and the essentially zero rewards for teachers' time devoted to other subjects, is what makes pay-for-performance *perilou*s. In activities as complex as education and learning, meeting the goals of comprehensive performance measurement and linking each measure to a reward structure, is fanciful.

It should come as no surprise that teaching-time reallocations, *to* math, English, and science, and *away from* the other unrewarded subjects, are not mere speculations. They have already taken place in U.S. schools in response to the federal focus on Math, English and later science testing. Evidence of schools' reallocating time was already surfacing by 2005, three years after NCLB dramatically elevated those subjects to schools' priorities among school districts that specified the amount of time to be spent on "reading/language arts" and on "math instruction," class time was reported to have been cut "somewhat" or "to a great extent" in a variety of other, unmeasured and unrewarded, subjects; 10 percent of school districts reported cuts in art and music time, and 27 percent in social studies time. [83]

The amount of time reallocated away from these subjects and toward math and English had been substantial. Among all of the school districts responding that they had increased class time devoted to the *tested* subjects in elementary schools, a follow-up study reported that the average time increase between 2002 and 2007 was 47 percent for English and 37 percent for math—a total reduction of 145 minutes a week devoted to other subjects, nearly a half hour a day.[84]

Like all survey responses, these are open to interpretation and verification; but they are consistent with the expected direction of responses to the greater rewards for better performance in math and English. At the same time, why the frequency of reported cutbacks differed among the *u*ntested subjects, such as physical education and social studies, deserves more attention, as do the magnitudes of the cutbacks and their long-run effects.

The consequences of time reductions in the non-tested, unrewarded, subjects are unknown and will remain unknown as long as there are no reliable standardized tests of performance in those areas. These issues are not new, they were not new in 1970 when President Nixon called for greater attention to *immeasurable* forms of schools' performance, nor were they new in 1860 Britain. Today we are still in search of standards, measures, and valuations of educational performance. Without consensus on them, there cannot be meaningful incentives to spur "performance." Educators have not even come close to consensus on how to quantify the many "standard" dimensions of education, let alone a school's success in advancing student "responsibility, wit, and humanity."

Conclusion: No Incentive System is Perfect, in Education or Anywhere Else, But We Can Do Better

Rewards, whether strong or weak, may serve important functions, but in U.S. education today the shortcomings of *strong* rewards in K-12 schooling have been largely ignored in the search for "greater efficiency."

[83] CEP, "NCLB: Narrowing the Curriculum?"
[84] McMurrer, "NCLB Year 5: Choices, Changes, and Challenges"

The pitfalls of strong rewards for measured performance that is multi-dimensional, costly-to-measure, and easy-to-game, are invariably troubling, the more the rewards for "good" or "improved" performance are tied to systematically mis-measured indicators. They do not justify, however, scrapping all P4P efforts. But neither is there a straightforward way to link strong, high-powered, rewards for teachers and school administrators to only the intended effects on students' school performance. Searching for one or two measures is futile, as is acting on the belief that we have found them, beguiling as that is. With today's understanding of measurement technologies there is reason to expect *in*efficiencies to increase if rewards for making the education system more "efficient" escalate.

Pay-for-Performance (P4P) in schools, as an idealized goal, has morphed into P4MP -- Pay-for-*Measured*-Performance. Once rewards are linked to particular measures of performance, P4P becomes P4MP, not pay for *true* performance. Among public sector industries, K-12 education has received the most effort to measure its performance quantitatively—by measuring students' scores on standardized tests in math and reading, and then rewarding high-performing schools, teachers, students, and administrators, while penalizing low-performers. Because performance is systematically, not randomly, *mi*smeasured, biased results continue to follow.

The potential for gaming the educational reward system is vast. The greater the rewards the more gaming will occur. Though *illegal* gaming—fraud—captures the most headlines when it is identified, the *legal* manipulations of rewards are more insidious, not attracting wide attention and not generally bringing heavy penalties even when exposed, but they do have unwanted effects.

The usefulness of these perspectives and lessons is not limited to K-12 schooling. This *framework* -- that considers the major stakeholders and the legality of methods each stakeholder might develop to game the reward system – can help direct attention to the broader gamut of potential forms of gaming across the education industry as well as other industries. It can also help to explain why there is great variation among government, nonprofit, and for-profit providers in their reliance on stronger rather and weaker rewards—and why weaker rewards with more multi-dimensional criteria can be the preferred path.

Bibliography

AASA [American Association of School Administrators].. "South Carolina, Texas Educators Receive 2014 AASA Women in School Leadership Awards", press release, Feb. 14, 2014. http://aasa.org/content.aspx?id=32130.

Alvarez and Marsal, *Final Report, District of Columbia Public Schools, Audit and Investigation*, Office of the State Superintendent of Education, January 26, 2018. https://osse.dc.gov/sites/default/files/dc/sites/osse/release_content/attachments/Report%20on%20DCPS%20Graduation%20and%20Attendance%20Outcomes%20-%20Alvarez%26Marsal.pdf

Aramark. "Atlanta School Leader Beverly Hall Named 2009 National Superintendent of the Year." Press release, February 20, 2009. http://www.aasa.org/content.aspx?id=1592 .

Aviv, Rachel. "Wrong Answer." *New Yorker,* July 21, 2014. http://www.newyorker.com/magazine/2014/07/21/wrong-answer

Baird-Remba, Rebecca. "Former 'Superintendent of the Year' Could Go to Prison for 45 Years." *Business Insider,* 2013. http://www.businessinsider.com/atlanta-cheating-indictment-beverly-hall-2013-4 .

Baker, Al. "Test Prep Endures in New York Schools, Despite Calls to Ease It." *New York Times*, April 30, 2014. http://www.nytimes.com/2014/05/01/education/test-prep-endures-in-new-york-schools-despite-calls-to-ease-it.html .

Banchero, Stephanie. "Indiana Drops Common Core." *Wall Street Journal,* August 20, 2014a. http://online.wsj.com/articles/indiana-drops-common-core-1395700559 .

Banchero, Stephanie. "States Look to Curb Standardized Testing." *Wall Street Journal,* February 28, 2014b. https://www.wsj.com/articles/SB10001424052702304071004579411433293118344

Beasley, David. "Jury Chosen in Cheating Trial of Former Atlanta Educators." *Reuters*, September 22, 2014. http://www.reuters.com/article/2014/09/22/us-usa-education-atlanta-idUSKCN0HH2UP20140922 .

Bidwell, Allie. "Florida School District opts Out of Opting Out," *U.S. News & World Report,* September 2, 2014. http://www.usnews.com/news/articles/2014/09/02/florida-school-district-retracts-historic-testing-opt-out-decision .

Bloom, Molly, and Bill Rankin. "Heard Most on Day 1 of APS Trial? 'Conspiracy' and 'Beverly Hall'." *Atlanta Journal-Constitution,* September 29, 2014. http://www.ajc.com/news/news/local/atlanta-school-cheating-trial-gets-underway/nhXmy/ .

Blow, Charles M. "The Do-Even-Less Congress." *New York Times*, August 4, 2014: A21.

Bowers, Michael J., Robert E. Wilson, and Richard L. Hyde. *Special Investigation into Test Tampering in Atlanta's Schools.* 3 vols. Atlanta, GA: Office of the Governor, Special Investigations, June 30, 2011. http://georgiataxcreditscholarship.org/our-motivation/atlanta-cheating-scandal .

Cardinale, Daniel J. "How to Get Kids to Class." *New York Times*, August 26, 2014: A23.

Carter, Chelsea J., with Dave Alsup, Joe Sutton, and Darrell Calhoun. "Grand Jury Indicts 35 in Georgia School Cheating Scandal." *CNN*, March 29, 2013. http://www.cnn.com/2013/03/29/us/georgia-cheating-scandal/ .

CBS/AP. "Philadelphia Principal, 4 Teachers Charged in Test-Cheating Scandal." *CBSNews.com*, May 8, 2014. http://www.cbsnews.com/news/philadelphia-principal-4-teachers-charged-in-test-cheating-scandal/ .

Center on Education Policy [CEP]. "NCLB: Narrowing the Curriculum?" *NCLB Policy Brief,* 3, July. Washington, DC: Center on Education Policy, 2005. http://www.cep-dc.org/displayDocument.cfm?DocumentID=239.

Chen, Grace. "Parents Refuse Common Core Testing," *Public School Review*, blog, updated January 7, 2021. https://www.publicschoolreview.com/blog/parents-refuse-common-core-testing

Common Core State Standards Initiative. "Standards in Your State," 2015. http://www.corestandards.org/standards-in-your-state/ .

Copeland, Larry. "School Cheating Scandal Shakes up Atlanta." *USA Today,* April 14, 2013. http://www.usatoday.com/story/news/nation/2013/04/13/atlanta-school-cheatring-race/2079327/ .

DeNardo, Mike. "With New Anti-Cheating Procedures, Philadelphia PSSA Test Scores Plummet." CBS Philly, September 21, 2012. http://philadelphia.cbslocal.com/2012/09/21/with-new-anti-cheating-procedures-philadelphia-pssa-test-scores-plummet/ .

FairTest. The National Center for Fair & Open Testing. "NCLB Boosts Temptation to Cheat." May 2004. http://fairtest.org/nclb-boosts-temptation-cheat .

Fausset, Richard. "Trial Opens in Atlanta School Cheating Scandal." *New York Times*, Sept. 30. 2014: A17.

Figlio, David N. "Testing, Crime and Punishment." *Journal of Public Economics* 90: 837–51, 2006.

Gladwell, Malcolm. *Outliers: The Story of Success*. New York: Little Brown & Company, 2008.

Golden, Daniel. "Student's Dream, Principal's Dread: The Test Not Taken." *Wall Street Journal,* December 24, 2002: A1, 6.

Graham, Kristen A., Martha Woodall, and Claudia Vargas. "Charges in 'Culture of Cheating" at Philadelphia School." *Philadelphia Inquirer,* May 10, 2014.

Gladwell, Malcolm. *Outliers: The Story of Success*. New York: Little Brown & Company, 2008.

Hanushek, Eric A., and Steven G. Rivkin. "Generalizations about Using Value-Added Measures of Teacher Quality." *American Economic Review: Papers & Proceedings* 100 (May 2010): 267–71.

Hernández, Javier C., and Al Baker. "A Tough New Test Spurs Protest and Tears." *New York Times,* Education §, April 19, 2013. http://www.nytimes.com/2013/04/19/education/common-core-testing-spurs-outrage-and-protest-among-parents.html .

Howard University, "Why Public School Teachers, Administrators Cheat," *AFRO News*, Feb 5, 2018. https://www.afro.com/public-school-teachers-administrators-cheat/

Jarvie, Jenny. "Atlanta School Cheating Trial Has Teachers Facing Prison." *Los Angeles Times,* September 6, 2014. http://www.latimes.com/nation/la-na-cheating-trial-20140907-story.html#page=1 .

Kahlenberg, Richard D., and Halley Potter. "The Original Charter School Vision." *New York Times*, August 31, 2014: SR12.

Koebler, Jason. "Educators Implicated in Atlanta Cheating Scandal." *U.S. News & World Report,* July 7, 2011. http://www.usnews.com/education/blogs/high-school-notes/2011/07/07/educators-implicated-in-atlanta-cheating-scandal .

Konnikova, Maria. "Youngest Kid, Smartest Kid?" *New Yorker,* September 19, 2013. http://www.newyorker.com/tech/elements/youngest-kid-smartest-kid .

Lattanzio, Vince. "Philadelphia Teachers, Principal Charged in Test Cheating Scandal." *NBC 10*, *Philadelphia*, May 8, 2014. http://www.nbcphiladelphia.com/news/local/Teachers-Principal-Charged-in-Philadelphia-School-Cheating-Scandal-258455781.html .

Layton, Lyndsey. “GAO: 40 States Have Suspected Cheating on K-12 Tests.” *Washington Post,* May 17, 2013. http://www.washingtonpost.com/local/education/gao-40-states-have-suspected-cheating-on-K-12-tests/2013/05/17/a366542c-bf1d-11e2-97d4-a479289a31f9_story.html .

LeBlanc,” More states ditch exams as high school grad requirements,” Associated Press, appearing in Boston Globe, November 17, 2024. https://www.boston.com/news/education/2024/11/17/more-states-ditch-exams-as-high-school-grad-requirements/

Martinez, Barbara, and Tom McGinty. “Students’ Regents Test Scores Bulge at 65.” *Wall Street Journal,* February 2, 2011. http://online.wsj.com/articles/SB10001424052748703445904576117793343465096.html .

McMurrer, Jennifer. “NCLB Year 5: Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era.” Washington, DC: Center on Education Policy, 2007. http://www.cep-dc.org/displayDocument.cfm?DocumentID=312 .

McWhirter, Cameron, and Stephanie Banchero. “Ex-Head of Atlanta Schools Indicted in Cheating Probe.” *Wall Street Journal,* March 29, 2013. http://online.wsj.com/articles/SB10001424127887323361804578391031169999780 .

McWhirter, Cameron, and Caroline Porter. “For School Tests, Measures to Detect Cheating Proliferate.” *Wall Street Journal,* September 26, 2014. http://online.wsj.com/articles/for-school-tests-measures-to-detect-cheating-proliferate-1411752291 .

Mead, Rebecca. “The Defiant Parents: Testing’s Discontents.” *New Yorker,* January 22, 2014. http://www.newyorker.com/news/daily-comment/the-defiant-parents-testings-discontents .

Medina, Jennifer. “State’s Exams Became Easier to Pass, Education Officials Say.” *New York Times*, July 20, 2010: A18.

New York Times Editorial Board, “Opting Out of Tests Isn’t the Answer.” Editorial, August 15, 2015: A18. https://www.nytimes.com/2015/08/15/opinion/opting-out-of-standardized-tests-isnt-the-answer.html

New York Times Editorial Board, “Caution and the Common Core.” Editorial. *New York Times*, May 28, 2013: A16. https://www.nytimes.com/2013/05/28/opinion/caution-and-the-common-core-state-education-standards.html

Nixon, Richard. “Special Message to the Congress on Education Reform.” March 3, 1970, in John Woolley and Gerhard Peters, *The American Presidency Project* [online]. Santa Barbara: University of California. http://www.presidency.ucsb.edu/ws/?pid=2895 .

No Child Left Behind Act of 2001 [NCLB]. Pub. L. No. 107–110; 115 Stat. 1425; 20 U.S.C. §6301 et seq.

Pennsylvania Department of Education. “2010–2011 PSSA and AYP Results,” 2011

Ravitch, Diane *The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education*, New York City, Basic Books, 2010.

Reese, William. *Testing Wars in the Public Schools*. Cambridge, MA: Harvard University Press, 2010.

Rich, Motoko. “States Given a Reprieve on Ratings of Teachers.” *New York Times*, August 22, 2014: A17. https://www.nytimes.com/2014/08/22/education/education-secretary-allows-reprieve-on-test-based-teacher-ratings.html

Rich, Motoko. “Home Schooling: More Pupils, Less Regulation.” *New York Times,* January 5, 2015: A1, 9.. https://www.nytimes.com/2015/01/05/education/home-schooling-more-pupils-less-regulation.html

Rich, Motoko, and John Hurdle. “Erased Answers on Tests in Philadelphia Lead to a Three-Year Cheating Scandal.” *New York Times*, January 24, 2014: A16.

Rothstein, Richard, with Rebecca Jacobsen, and Tamara Wilder. *Grading Education: Getting Accountability Right*. Washington, DC: Economic Policy Institute, and New York: Teachers College Press, 2008.

Schemo, Diana Jean. “Ninth Grade Key to Success, but Reasons Are Debated.” *New York Times*, January 18, 2004. http://www.nytimes.com/2004/01/18/education/18NINT.html .

Severson, Kim. “A Scandal of Cheating, and a Fall from Grace.” *New York Times*, September 7, 2011. http://www.nytimes.com/2011/09/08/us/08hall.html

Simkin, John. “Robert Lowe.” Spartacus Educational. http://spartacus-educational.com/EDlowe.htm , 2020.

U.S. Department of Education, Office of Special Education Programs. *Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act, 2013.*

Weisbrod, Burton A. “Toward a Theory of the Voluntary Nonprofit Sector in a Three-Sector Economy.” In *Altruism, Morality and Economic Theory,* edited by Edmund S. Phelps, 171–95. New York: Russell Sage, 1975.

Weisbrod, “Why Strong Performance Rewards in Government and Nonprofit Programs Don’t Work -- A Social Heisenberg Principle”, paper 1 in The Perils of Pay for Performance of Public Service Industries, Institute for Policy Research, Northwestern University, Working Paper Series, WP25-01.

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

Wisconsin Education Alliance Council [WEAC]. “What Do We Know about Merit Pay?” Research Brief 20. Madison, WI: WEAC, 2011. http://bloggingblue.com/2013/05/weac-the-problem-with-merit-pay