# SURVEILLANCE AND EDUCATIONAL TESTING: NO CHILD LEFT BEHIND AND THE REMAKING OF AMERICAN SCHOOLS

John Gilliom


## ABSTRACT

*Educational testing launched under "No Child Left Behind" (NCLB) brings unprecedented levels of surveillance to public education in the U.S. The testing regime is moving American pedagogy away from types of teaching which are either politically disfavored or not easily tested. The impact of NCLB will be strongest in lower-income schools which fare poorly on such tests; these schools can expect to see sanctions, shaming, and a concomitant departure of committed families and teachers. The reshaping of American education wrought by NCLB compels us to reimagine mass surveillance as not primarily a means of watching the world, but as expressions of power capable of effecting significant changes in institutions and behaviors.*


**Surveillance and Governance: Crime Control and Beyond**
**Sociology of Crime, Law and Deviance, Volume 10, 305–325**

**ISSN: 1521-6136/doi:10.1016/S1521-6136(07)00214-X**

## INTRODUCTION

> But who will write the more general, more fluid, but also more determinant history of the "examination" – its rituals, its methods, its characters and their roles, its plays of questions and answers, its system of marking and classification? For in this slender technique are to be found a whole domain of knowledge, a whole type of power.
>
> – Foucault (1979, p. 185)

In 2002, President George W. Bush signed a reauthorization of national education law that has come to be known as No Child Left Behind (NCLB). In so doing, he unleashed what is probably the most ambitious surveillance program in the history of the nation. Under the new program, tens of millions of Americans, in every community and state, are subject to unprecedented frequencies and degrees of state monitoring. The educational testing regime set forth under NCLB is a transformative moment in the history of not just education, but government surveillance itself. It thus offers a rare opportunity to explore surveillance and the legal and political conflict over its impact and implementation.

Perhaps the most fundamental lesson to be learned is one that has been encountered in numerous prior studies of the politics of surveillance: even the most technocratic forms of surveillance tend to extend and enforce extant patterns of race and class bias, extant moralist and criminological agenda, and extant assumptions about subject populations. In *Surveillance, Privacy, and the Law: Employee Drug Testing and the Politics of Social Control* (Gilliom, 1994), I found that the employee drug testing initiative of the 1980s was embedded within the broader anti-union and anti-labor politics of the Reagan era. The testing movement could only be understood as a political–cultural bid to infuse workplace politics and ongoing labor–management struggles with the broad patina of the so-called War on Drugs. Evidence that workplace drug use was relatively minor problem outside of a few industry sectors, that drug testing was an ineffective response, and that the drug most likely to be detected, marijuana, was the least worrisome, fell by the wayside under an onslaught of pro-testing propaganda and media cooperation.

In *Overseers of the Poor* (Gilliom, 2001), a study of the computerized surveillance of welfare clients, a major surveillance initiative was launched to deal with relatively low levels of fraud and abuse in state welfare systems. Accompanying the launch of the new programs were major media campaigns highlighting anecdotal stories of welfare fraud and portraying the poor as abusers and cheats in need of close monitoring to protect the taxpayers' money. The implementation of the system itself brought salvoes

of shame upon welfare clients with the constant trumpeting of the assumption that they were cheats and frauds adding to the already extreme social stigma of poverty in America. In welfare surveillance, finally, the optic and metrics of the system, created a bizarre state-centered view of poor Americans, a depiction that belied a complex reality with partial figments of the bureaucratic imagination.

Each of these studies joined a body of research and analysis finding that government surveillance and information initiatives are almost assured to miscast and misperceive their subjects in ways which not only express the political, cultural, and technological constraints on policy, but more importantly, render a body of data and knowledge that ensures poor policy and a failure of statecraft (Scott, 1998). In the present examination of the optics and implementation of surveillance of America's teachers, schools, and schoolchildren, we see another chapter in the unfolding story of the failure of surveillance in contemporary governance – budget limitations, cultural biases, technological shortcomings, and political preferences combine to render a system that will see what it is able to see and see it poorly at that.

Surveillance has, of course, always been a part of formal education. From simple things like quizzes, tests, assignments, and attendance records, classroom teachers monitor and assess the work of their students. Their goals are multifaceted. Teachers want to see if students are absorbing and retaining information, learning to use new skills, and developing the capacity for critical thinking. Teachers also use testing as a means to compel students to do the reading and coursework, to take the course seriously, and, inevitably and sometimes unconsciously, as a way of flexing bureaucratic muscle in the politics of the student–teacher relationship. Anyone reading this has at some point been a student and, probably, a teacher; we all know from those experiences that surveillance, understood as observation, monitoring, and evaluation, is a big part of the game.[1]

The new surveillance manifest in the emergent regime of standardized testing is a fundamentally different thing than the older regime of quizzes and blue books. It is a fundamentally different thing in terms of sheer size and scale; in terms of the nature and targets of sanction; in terms of the expressions of political will and the centralization of power; and in terms of the standardization of American education. The exertion of power in NCLB is evident enough that Frederick M. Hess uses the term "coercive accountability" to describe the policy and Jones et al. use the term "measurement driven reform."[2] Each of these phrases reminds us that the testing apparatus established under NCLB is not some sort of neutral or

inconsequential assessment: there are policy goals, outcomes, and implications from any system of social assessment like this. And there are also, as we shall see, critically important implications tied to the nature, limitations, and use of the mechanisms of observation. In the end, this essay will argue that the testing movement achieves such a radical redrawing or reimaging of American education that it compels us to confront the question of whether the term "surveillance" and its idea of simply watching really captures the full magnitude of this initiative and its many companions. Massive new systems of social surveillance have come to be such powerful bodies of communication, depiction, and social organization that the old language of observation is obsolete.

## NO CHILD LEFT BEHIND: THE BASICS

> The purpose of this subchapter is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments. This purpose can be accomplished by ... ensuring that high-quality academic assessments, accountability systems, teacher preparation and training, curriculum, and instructional materials are aligned with challenging state academic standards so that students, teachers, parents, and administrators can measure progress against common expectations for student academic achievement ...
>
> – *U.S. Code: Title 20: Section 6301*

NCLB is a reauthorization and revision of the 1965 Elementary and Secondary Education Act (ESEA). The carrot and stick behind NCLB is a budgetary item called Title I which started as the 1965 Act's effort to steer federal money to disadvantaged children. It is the single largest source of federal money for education (almost $12 billion in 2003) with aid going to 90 percent of school districts in the nation (Rudalevige, 2003, p. 25). It is the threat of losing this money that serves as the leverage to bring states into compliance with federal guidelines. In short, no states have been *commanded* to obey NCLB, but nor has any state been willing and able to forego the federal dollars attached to the educational surveillance initiative in NCLB.

NCLB is a massive initiative, but for our purposes we can focus on the following key features:

- States must establish a standards-based curriculum and design tests to assess mastery of that curriculum.
- Annual tests are required in reading and math for all students in grades 3–8 and once in high school; science testing is mandated after 2007–2008.

- Test scores must be reported for schools and for significant subgroups within schools, including ESL students, disabled students, with population breakdowns by race and family income.
- Test scores for schools and subgroups are assessed to determine whether a school has made "adequate yearly progress" (AYP); failure to achieve AYP in multiple years earns escalating sanctions. While only about half of the nation's schools receive Title I funding, NCLB requires all schools to meet the standards. The actual Title I schools (meaning the poorer ones) face the strictest public accounting and sanctions of NCLB; after two to four years of inadequate progress they face such measures as state takeover, staff replacement, or conversion to charter school.
- States must have *all* children up to proficiency by 2014 and make measurable progress toward that goal during the years leading up to 2014. That progress must be evident in each identified subgroup.
- "States are free to determine their own standards, to create their own tests, and to determine for themselves the scores that individual students must receive in order to be deemed 'proficient.'" (Ryan, 2004, pp. 941–942).[3]
- NCLB also mandates that the National Assessment of Education Progress (NAEP) examinations in math and reading be given every two years to fourth and eighth graders. NAEP uses a random sample testing method to create a composite of a state's progress on a national metric of skills and content. It is frequently used as that national standard for comparisons among states and nations.

Much of what emerged in the 2001 NCLB had been circulating in the educational policy world for the last couple of decades. Many observers point to the Reagan era study *A Nation at Risk* as the most visible starting point for the modern push for standards-based educational assessment.[4] Momentum for performance standards grew during the Bush I and Clinton eras, with Bill Clinton (as both president and governor) playing a key role in the push toward setting education outcome goals. Indeed the 1994 reauthorization of the 1965 Elementary and Secondary Education Act mandated standards, testing and the AYP framework to such an extent that it "fundamentally changed the nature of Title I. Instead of providing funds to support remedial instruction for disadvantaged students, Title I funds now had to be used to create standards for all students" (Ryan, 2004, pp. 938–939). But the laws of the 1990s, though they called for state standards and testing, left enough room for flexibility and interpretation that "work progressed very slowly" (Rudalevige, 2003, p. 29). During this period a rough battle line emerged in national politics with Republicans

being opposed to a national curriculum on state's right grounds and Democrats being opposed to a regime of standards-based tests (Rudalevige, 2003, p. 30).

With conservative Democrats joining Republicans on a number of fronts, a series of acts in the very late 1990s put into partial action most of the key elements of NCLB. Numerical goals, AYP, public report cards, state-set standards, and the menu of sanctions were all developed during this period. Meanwhile in Texas, tycoon Ross Perot had led the state to a system including annual testing and the disaggregation of school data to compel ethnic and economic subgroup assessments (Hess, 2003, p. 68). When the incoming Bush administration laid out the blueprint for what would become NCLB, one aide to Senator Lieberman was able to say that Bush "essentially plagiarized our plan" (Rudalevige, 2003, p. 36). But that is not quite right. What emerged in NCLB was a convergence of some longstanding Washington ideas and some of the key Texas policies developed in the 1990s. In an unprecedented convergence, this author and the Bush administration agree that the testing and accountability measures were the most important elements of NCLB. In the logrolling and compromise that resulted in NCLB, the Bush team even gave up their interests in school vouchers, and faced the wrath of cultural conservatives and the loss of many Republican votes in the Congress, in order to preserve the testing mandate.

The following analysis of the surveillance mechanisms in NCLB addresses several main points. Namely, that the methods of assessment are designed in such a way as to reduce their effectiveness for authentic educational outcome assessment; that the key metric AYP promotes lower standards and racial and economic segregation; and that there are fundamental fallacies at the heart of the idea of a single measurement of academic performance. Subsequently, the chapter explores the ways in which NCLB testing protocols are reshaping the curricula of U.S. schools, particularly those with concentrations of low income and minority students.

## OPTICAL ILLUSIONS

As a tool of observation and assessment, the testing established under NCLB is profoundly flawed. Ignoring all of the questions of politics, values, and power struggles to artificially isolate the simple process of assessment, the distortions and limitations are enormously important.

## *The Optics of Mediocrity*

The stated goal of NCLB is to improve learning. In pursuit of this, NCLB launched an ambitious program of frequent high-stakes testing, but there is considerable research indicating that state level high-stakes testing has no discernible outcome on learning. At first glance, this may seem counter-intuitive because scores have, in general, been rising on state level tests since they became popular in the 1990s and then expanded under NCLB in the early 2000s. Therefore, learning must be improving. And learning may indeed be improving, but it may just be learning about how to take the statewide tests as teachers and administrators become more familiar with the content and question design and are able to train their students to more successfully fill in the forms.[5] In this section, we begin by reviewing data indicating that more testing does not necessarily mean more learning and then examine the metrics and finances of NCLB testing to see how improving state test scores may be relatively easy and that the improvements achieved may not indicate improvements in educational outcomes.

When skeptical researchers check up on the efficacy of statewide tests by examining how students do on other examinations, they find little to suggest that rigorous state testing is associated with student improvement:

> Analyses of scores and participation rates for the NAEP, ACT, SAT and AP tests suggest that there is inadequate evidence to support the proposition that high-stakes tests and high school graduation exams increase student achievement. The data presented in this study suggest that after the implementation of high stakes tests, nothing much happens … The data presented in this study also suggest, however, that after the implementation of high school graduation exams, academic achievement apparently decreases … ACT, SAT and AP scores decline. Indeed, on balance, the analyses suggest that high-stakes tests and high school graduation exams may tend to inhibit the academic achievement of students, not foster academic growth. (Amrein & Berliner, 2002, pp. 57–58)

So even on one of the most basic and widely agreed upon goals of NCLB, we may actually get exactly the opposite of what was promised.

One clear area of impact, intended or not, is that the educational curricula of schools across the nation are being changed not just to meet the newly required state *content* guidelines but to teach students with *methods* that are most conducive to success on the testing instrument. We know that content is shifting more toward the tested areas of math, reading, and, increasingly, science, with corresponding cuts in arts, music, and other untested areas, particularly in high need areas. But what is also occurring is a shift of education within areas like math and English as teachers are compelled to move away from broader, more theoretically rich approaches to forms of

instruction that emphasize the smaller particle and tactics that can be affordably tested. Furthermore, as a state's testing style and content become known, teachers can orient to the specific content and question styles of given tests.

Such outcomes point to one of the key vulnerabilities of standardized educational testing as a mechanism of surveillance. As testing expert W. James Popham (2005) explains, any test measures only a tiny subset of relative knowledge. An ideal test, on "everything," would take too long to complete, grade, and process, so we pick a relatively tiny subset, or sample, of materials. The fine art or science of designing that subset is the work of experts known as psychometricians. As states are unable to find or afford competent psychometricians, or overwhelm the ones they have, the test questions and question structures see more and more repetition from year to year and the tests become easier and easier to game. Teachers learn the pattern of the test, old tests are used for training, private sector preparation books enter the market; the end result is that the test itself becomes the object of study and the successful negotiation of its terms the goal of pedagogy. The training and gaming allows school districts to show improvement, but it is improvement at taking the tests, not necessarily improvement on mastering the universe of knowledge and skill that the tests seek to assess. For authentic proponents of testing as a real measure of educational progress, such outcomes are deeply problematic because they render the tests far less meaningful as tools of assessment. As Popham explains, if that tiny sample of knowledge that is to be tested becomes known in advance or over time, the entire logic of the assessment tool collapses.

The problems are compounded by the aforementioned fact that standardized educational testing in the U.S. has to be done on the cheap. The explosion of testing in the wake of NCLB was so enormous that by 2006 there was actually a test question crisis in America. As of this writing, the dramatic expansion of standardized educational testing has created a severe shortage of psychometricians and a true problem in the industry's capacity to make competently designed test questions (Toch, 2006). Part of the problem is created by the quintessentially American federalist approach to NCLB. Under the compromise brokered in the Congress, there is a national educational accountability program manifest in the demand that all children have annual tests in grades 3–8 and one more to graduate from high school. But there was also a nod to federalism in allowing each state to set its own educational standards and design its own exams. With this nod, the hope of a technically competent testing regime was destroyed: there are simply not enough qualified personnel to design a testing regime which includes not just several grade levels in each of 50 states with their

50 curricula, but the many other testing programs under state minimal competency laws *and* the classics such as the SAT, ACT, LSAT, MCAT, etc.[6] Furthermore, as has been widely noted, NCLB provided little in the way of funding to support the new testing programs. According to the U.S. General Accounting Office, multiple choice tests run about one to two dollars per student while more comprehensive performance assessments range from 35 up to 70 dollars per student (Jones, Jones, & Hargrove, 2003, pp. 16–17). This has created a budget situation in which the only feasible forms of testing are the most simple multiple choice tests – the slightest nod to an essay format or creative student responses puts the cost of evaluating the tests through the ceiling. Under economic and time pressures, many states have opted for the cheapest and most easily scored types of test questions – multiple choice questions measuring basic rote knowledge.

One of the key implications of this helps us frame and understand the recent evidence of test score improvement that we see in many school districts. With a limited capacity to create test questions that are either truly creative or well-designed, districts and teachers are readily able to "teach to the test" – essentially training students to the sorts of questions that can be expected and, therefore, over time, creating an image of improvement. Such training need not be in the form of actually teaching the direct content of the test, though this certainly occurs. Teachers can, for example, spend time teaching what they call "look-alikes." Once it is known that tests will structure the multiplication and division questions in certain formats, teachers can orient their instruction to that format. There is no necessary improvement in instruction, in fact – as we will soon see – it may be less effective. But such practices do lead to better measurements (Jones et al., 2003, p. 66). Indeed, there was a major scandal for the state of Texas when a RAND corporation analysis found disparities between student performance on state level testing and the national standards exam. The conclusion was that teachers in Texas had figured out enough about the statewide test that they were able to prep their students, who showed improvement over the years. There was also evidence that school district administrators had cheated. Once the children of Texas faced the differently formatted NAEP, the evidence of improvement vanished.

### *Absolute Criteria and the Reshaping of American Schools*

A critically important choice was made in designing the surveillance procedures at the heart of NCLB to work with an absolute criteria scale

of assessment rather than a value-added or rate-of-improvement scale. Under the current absolute criteria approach, if Nebraska sets its statewide curriculum standards to the expectation that every third grader can do long division to the fourth decimal point, then every school in the state must have a certain percentage of their students meet that standard to avoid being designated as failing. Yet under a value-added approach, if a school had, say, been able to get a given percentage of its third graders to do long division a certain measure better than they could do it in the second grade, the school would be judged successful even if the students could not make it out to the fourth decimal point. The value-added approach is more flexibly applied to the different contexts in which school and teachers function and has the potential to add accountability while avoiding a number of what have been called "perverse incentives" (Ryan, 2004) in the absolute standards approach.

> To appreciate the promise of value-added assessments, it is important to revisit why (the absolute criteria approach to) AYP is a relatively useless measure of school quality. As mentioned above, student performance is the product of a number of factors, some of which schools can control, others of which are beyond a school's ability to influence. A student's score on a standardized test is the result of both school and teacher inputs, as well as a host of exogenous factors, including innate ability, socioeconomic status, parental involvement, community stability, and peers. Because of the influence of these exogenous factors, looking to whether students in a school hit a uniform benchmark of achievement-the current approach to measuring AYP–actually tells us very little about the quality of the school itself. (Ryan, 2004, pp. 978–979)

For example, schools with relatively advantaged students typically post better test scores than those with relatively disadvantaged students. But it does not follow that the former schools are better at educating students than the latter; the scores may simply reflect the fact that the former school has students who take tests better than those at the latter. It is a well-known truism in the testing business that most assessment tests largely assess the socioeconomic background, or "social capital" of the students taking the test – known as the "Volvo effect" because, as Jones et al. (2003) summarize, "simply court the number of Volvos, BMWs, or Mercedes owned by the family and you have a good indicator of how well the child will perform on standardized test" (p. 118). Testing critic Alfie Kohn gives several examples of the Volvo effect at work:

> A study of math scores on the 1992 National Assessment of Educational Progress found that the combination of four variables unrelated to instruction (number of parents living at home, parents' educational background, type of community [e.g., "disadvantaged urban," "extreme rural"], and state poverty rate) explained a whopping 89 percent of the

> differences in state scores. In fact, one of those variables, the number of students who had one parent living at home, accounted for 71 percent of the variance all by itself … The same pattern holds within states. In Massachusetts, five factors explained 90 percent of the variance in scores on the Massachusetts Comprehensive Assessment System (MCAS) exam, leading a researcher to conclude that students' performance "has almost everything to do with parental socioeconomic backgrounds and less to do with teachers, curricula, or what the children learned in the classroom." … Another study looked just at the poverty level in each of 593 districts in Ohio and found a .80 correlation with 1997 scores on that state's proficiency test, meaning that this measure alone explained nearly two-thirds of the differences in test results … Even a quick look at the grades given to Florida schools under that state's new rating system found that "no school where less than 10% of the students qualify for free lunch scored below a C, and no school where more than 80% of the students qualify scored above a C. (Kohn, 2001)

The absolute standards approach that is currently used in the U.S. asserts that an overwhelming percentage of students must achieve certain statewide benchmarks by certain dates and at the exact same as every other child in the state. The implications of this choice are enormous. As Ryan (2004) explains, schools that do not show AYP are marked as failures and face media attention, financial sanctions, and professional shame. Staff may be let go and state agencies may take over. Parents with the financial wherewithal may move away while any parents with children in such schools are theoretically able to move their children to a better school in the district.[7] We should expect these effects to be more intense for schools with a significant number of low income and minority students. This is because, as discussed elsewhere, there is a distinct class and race bias to standardized testing outcomes, but also because such schools are most likely to be subject to the highest stakes testing (Amrein & Berliner, 2002, pp. 12–13). Schools with high numbers of lower income and minority students are more likely to be direct recipients of specific Title I funding, meaning that they face the strictest sanctions for poor performance. Higher hurdles, higher stakes.

This leads to what Ryan (2004) names the "perverse incentives" of NCLB. Working from the well-founded premise that success on absolute criteria test scores is primarily a measure of the extent to which a school is white and affluent, Ryan explains how weaker schools will slide further behind as quality teachers and educationally committed families bail out. Furthermore, racial and socioeconomic segregation will increase as affluent families flee for the better schools and the better schools begin to exercise self-protective measures to avoid taking weaker students. Schools also have an interest in urging low-performing students to drop out and to avoid accepting students who appear to be at risk of low performance. The results, argues Ryan (2004), are predictable: a further cementing of the class and

race gaps in American education as a poorly designed measure of school quality becomes the vernacular in American education.

Furthermore, Ryan argues the absolute standards approach to test design fuels a race to the bottom in the setting of state standards. Because all of the children in the state must pass the same nominal threshold at the same time, educational standards and testing criteria must be pushed down to a level where it is feasible for a significant number of schools and districts to pass. State leaders and education agencies will find it politically unacceptable to have an enormous number of schools failing the test and so they will be compelled to lower the bar until enough are able to make it over.

> Indeed, some already have. Louisiana, Colorado, Connecticut, and Texas have all tinkered with their scoring systems in order to increase the number of students who will be deemed proficient for purposes of the NCLBA. In Louisiana, for example, passing scores had been divided into three categories: basic, proficient, and advanced. Last year, only 17% of eighth graders scored at the proficient or advanced level on an English test, while 31% scored at the basic category; in math, only 5% were advanced or proficient while 37% scored at the basic level. So what did Louisiana do? It deemed those who scored at the basic level "proficient" for purposes of the NCLBA. Similarly, Colorado and Connecticut have redefined categories of scores, making it easier for students to reach the newly dubbed "proficient" level. And the Texas State Board of Education, after a field trial of state tests, lowered the number of questions students must answer correctly in order to be considered proficient on the third-grade reading test. (Ryan, 2004, p. 948)

The end results of the assessment choices built into the NCLBA is that "while the Act is supposed to raise achievement across all schools, it creates incentives for states to lower academic standards. Second, while the Act is supposed to close the achievement gap, it creates incentives to increase segregation by class and race and to push low-performing students out of school entirely, which will make it even more difficult for disadvantaged students to catch up to their more affluent peers. Finally, while the Act is supposed to bring talented teachers to every classroom, it may deter some from teaching altogether and divert others away from the most challenging classrooms, where they are needed the most. In short, although the Act is supposed to promote excellence and equity, it may work against both" (Ryan, 2004, p. 934).

### *A New Curriculum*

Several studies find evidence of a notable shift in teaching priorities in response to the NCLB testing. A national study of teachers found that

"A large majority of teachers felt that there is so much pressure for high scores on the state-mandated tests that they had little time to teach anything not covered on the test" (Pedulla et al., 2003, p. 2). Another study of teachers found that "teachers reported that after the implementation of the testing program, they spent substantially more time teaching the tested subjects of mathematics, reading, and writing and less time teaching science, social studies, the arts, and physical education and health … This narrowing of the curriculum has been reported in virtually every state where there is high-stakes testing of only a few subjects" (Jones et al., 2003, pp. 29–30).

Jones and his colleagues report that:

> testing sharply defines the knowledge and skill that students will learn … Prior to high stakes testing, teachers made the decision about what to teach within a broad framework of topics. Testing, however, not only defines what will be taught, but also defines the context of the knowledge. Whereas teachers may have previously embedded instruction in integrated units or taught concepts across multiple grades, testing necessitates that topics be taught in ways that can be assessed through discrete items on written tests given at very specific point of time. (Jones et al., 2003, p. 26)

As the Pedulla study team found, "Across all types of testing programs, teachers reported increased time spent on subject areas that are tested and less time on areas not tested. They also reported that testing has influenced time spent using a variety of instructional methods such as whole-group instruction, individual seat-work, cooperative learning, and using problems similar to those on the test" (Pedulla et al., 2003, p. 4).[8] They also found that teachers in states with particularly high-stakes testing programs are more apt to "engage in test preparation earlier in the school year; spend more time on such initiatives; target special groups of students for more intense preparation; use materials that closely resemble the test; use commercially or state-developed test-specific preparation materials; use released items from the state test; and try to motivate their students to do well on the state test" (2003, p. 5).

The primary educational result of the NCLB optics of surveillance may be a narrowing of the American K-12 curriculum to an outsized focus on training students to make correct choices between simple answers to questions in the three Rs: readin', 'ritin', and 'rithmetic. In many states, it will be just two of the Rs: there will be no 'ritin' because it is too expensive to score (Toch, 2006). These effects will be magnified in working class and minority schools because it is here where the tests are hardest to pass and the stakes are most dire. NCLB thus absorbs, cements, and advances our longstanding system of class and race stratification in the American education system: The surveillance optics and metrics are set in ways that

largely give upper income populations an easy pass while giving lower income and ESL populations almost insurmountable hurdles. The strongest sanctions for inevitable failure are reserved for those schools that receive Title I funds which are, by definition, the lower income school populations; those schools must adapt their curriculum in order to get over the testing hurdle; the result is that lower class schools focus on testing content and practices while more affluent school systems put up with the minor nuisance of a week or so lost to filling in some bubbles.

This section began with the argument that surveillance is not really surveillance – it is not mere watching, but must, rather, be understood as a form of creative depiction and world-making. Here, we have seen how true that is and that it goes far beyond the merely symbolic terms conveyed by the idea of depiction. In an act of veritable world-making, the standardized testing movement compels teachers and students across the nation to shape the content of the educational system to match the measures of the tests. The compulsion is far stronger in lower income schools than affluent and, therefore, stronger for people of color than it is for whites. The result can only be a strengthening of the extant class biases in American education, with poorer schools being pushed to the fragmented, technical archives of information that are so readily assessed.

## FUNDAMENTAL FALLACIES

Steven J. Gould's *The Mismeasure of Man* (Gould, 1996) remains the most essential and philosophically rigorous exploration of the fallacies of the testing movement. While Gould keeps his primary focus on general intelligence testing, three of the central testing fallacies are relevant for the types of standard assessment and proficiency ratings discussed here:

*Reductionism*. Among the most important fallacies is the error of thinking that the intelligence or proficiency of an individual, let alone a school, can be expressed as a single number or scale item. Simple rating systems necessarily belie the complexity, mutability, and incomparability of human and institutional qualities. A rating system that stamps an entire school with the label "Academic Emergency" or "Outstanding" is little more than a compilation of myths and exclusions that mocks the real and difficult process of assessing institutions.

*Reification*. Another fallacy is to treat a reductionist creation as a real thing; to invent concepts like "intelligence" or "school quality" and then speak of them as if it were a real thing, in a fixed location. Any attempt to

measure and codify necessarily creates reified illusions of certainty and reified partial snapshots of complex realities. When we then speak of these things as knowable and manageable entities – "let's bring our quality up" – we have given false and misguiding life to a figment of our cultural imagination.

*Ranking*. This occurs when we hierarchically sort individuals or schools on the basis of the reductionist reifications we construct. Any state will have a variety of schools – some are poor, some are moderately affluent, some are rich; some serve rural areas, some serve university towns, some focus on a broad and tracked curriculum, and some offer a more narrow and egalitarian set of courses. To take all these unique and different programs and institutions and somehow say that one is superior to another belies the complex and multifaceted nature of not just the institutions, but their locations, our values, and our metrics.

In *Overseers of the Poor* (Gilliom, 2001), I argued that one of the great fallacies and errors of financial surveillance in welfare administration was the artificial reduction of the multitudinous and diverse people known as "the poor" to a simple set of figures and statuses. The true and important multidimensionality, variance, and depth of families and individuals were subjugated for a simple terminology that the state's rule system established and managed. Using these terminological reifications, families were then sorted and ranked according to the complex rules of eligibility. The result is a false and incomplete fiction that takes on all the immense power of state action.

My concern about this transcends aesthetic mourning for the lost richness of humanity. As James C. Scott demonstrates in *Seeing Like a State* (Scott, 1998), the rational modernist government is doomed to failure at public policy interventions because its optics, or ways of seeing, necessarily simplify, reify, and reorder from a state-centric perspective. The resulting information and informational regime will invariably err because successful policy interventions require the sorts of complicated local knowledge and wisdom that the bureaucratic state simply cannot see. As I summarized in *Overseers of the Poor, Seeing Like a State*

> suggests that the power of surveillance is often an almost bumbling power which miscasts the world and its inhabitants, overlooks essential points of information, and helps generate the seeds of its own resistance through its ongoing misreadings of local knowledge. Scott argues that modern states must produce knowledge and information to guide their various missions of social intervention and design. To do so, they must both simplify a complex social reality and rewrite the terms of that reality to fit the terms of the intervention: a mass of people becomes a list, and last names and even

> street addresses are introduced as ways of organizing and knowing the population. A wilderness becomes a forest with full analysis of species and harvesting schedules. Scott argues that modern statecraft requires these systems of knowledge and that, further, it is these very systems of knowledge which doom statecraft to failure. The failure of statecraft is virtually guaranteed, he argues, because the systematic state knowledge necessarily omits or overruns the sorts of local and varied knowledge and practices that are inherent to any setting. Since these local forms of knowledge would be essential to the success of state planning, their omission essentially guarantees failure, as well as conflict and resistance from subjected peoples. (Gilliom, 2001, p. 131)

As an example, let me point to a unique high school in the Appalachian region of Southeast Ohio. It is the consolidated high school for one of the poorest, most rural, and geographically dispersed school districts in the state. The academic performance of its students is low and the teachers are poorly paid. The local taxpayers persistently refuse to pass levies while the state legislature perennially balks at reforming the property tax system of educational finance in Ohio even as the State Supreme Court has repeatedly found it unconstitutional. In the NCLB mandated rating system, the high school has only recently emerged from Academic Emergency status to hover between Continuous Improvement and Effective – low to middling scores. Most of its tests scores now run very close to or pass the state minimums, though science and social studies continue well below the norm (see: http://www.ode.state.oh.us/reportcardfiles/2006-2007/DIST/045914.pdf).

From the snapshot created by NCLB, it would be impossible to learn that the school is led by an award-winning educator or that its program of democratic education has made the school a national beacon of educational innovation. As summarized by the Center for Secondary School Redesign:

> In spite of the challenges faced in this region, the school has received numerous awards under Dr. Wood's leadership including an Ohio's Best Award for the school's internship program, designation as a First Amendment School by ASCD and the Freedom Forum's First Amendment Project for the school's work in promoting active democratic citizenship, and being named one of the first five Coalition of Essential Schools 'Mentor Schools' in conjunction with the work of the Gates Foundation. (http://www.cssr.us/keynotes.htm)

This high school represents a unique and intriguing social and educational experiment which appears to be working, but all of these dynamics are invisible to the state profiles created by the NCLB surveillance technology. As the school modifies its curriculum to satisfy the commands of the state curriculum and surveillance program, there may well be catastrophic consequences for the innovations currently underway. As Jones et al. (1993) show, systems of high-stakes testing not only modify curricula, absorb time,

and reduce resources for anything outside of the tested subject areas, they create pressures that reduce staff morale, push teachers out of the profession, and suppress teaching techniques that stray from the necessary pedagogy of the testing regime. That pedagogy, they show, is one that focuses on the rote acquisition of basic skills such that students can readily respond to simple, discrete, multiple choice questions in a known format. Such programs at Federal Hocking High School, which focus on democratic values, first amendment freedoms, and engagement through internships, face a hard time under the new optics of educational assessment.

## CONCLUSION: EDUCATIONAL TESTING, EDUCATIONAL SURVEILLANCE

Each year, some 50 million standardized tests are administered to the children of America. The results are used to develop files on individuals, assess the work of teachers and school leaders, rate schools and districts, and, if current laws are followed, deliver severe financial and political penalties to schools that fail to "measure up." The less formal effects are equally important. The testing movement is reshaping the American school curriculum, centralizing control over educational decisions, transforming pedagogy, and shifting billions of dollars of funding into the testing industry coffers. NCLB is, arguably, the greatest single national event in the history of American education.

And it is also one of the greatest expansions of mass surveillance in American history. We may initially balk at thinking of educational testing as surveillance because when we think of surveillance the mind first turns to things like eavesdropping, spy satellites, and phone taps. But a brief reflection on the idea and processes of surveillance makes it clear that educational testing falls into this category. "Surveillance" is derived from French, with a rough translation being "to watch (veil) from above (sur)." Surveillance has been widely studied as a form of management, political domination, and social control (see Lyon, 2007). The field rightly covers a host of policies and technologies from obvious practices like the increasing use of closed circuit cameras and national identification cards to the less popularly recognized but no less important surveillance practices like insurance scoring, credit reporting, and computer monitoring of social welfare clients (Lyon, 2007; Monohan, 2006; Haggerty & Ericson, 2006; Gilliom, 2001).

Surveillance programs of varying stripes gather information through watching, measuring, and monitoring around some set of norms, rules, or expectations. That information is used, in varying ways, as a means of control. The control agenda could be nearly anything. Reducing crime, illegal drug use, or illegal immigration; reducing fraud and error in the management of social welfare and healthcare systems; reducing traffic infractions such as speeding and stop-light violations; reducing the risk of violence in air transportation, schools, and public buildings; fighting the emergence and transmission of disease; building a successful college or law school cohort; and enhancing public education. The element that unites all of these areas is that the primary method of action is observation and assessment.

In some ways, it is unfortunate that "surveillance" has come to be the accepted term for these practices, for its specific connotation of watching is an inappropriate meaning for the policies and technologies at hand. It is critical that we should not lose sight of the fact that surveillance is not *really* best understood as an act of watching – it is often far more importantly understood as an act of depiction: a creative rendering of an impression of a part of the social environment. I raise this because the idea of watching implies a simple observation of a given object. It is sort of what you might get from a junior high school level understanding of what journalists do – "we report, you decide." As if there were no decisions going into the complicated process of observing, interviewing, ordering, narrating, editing, and pitching. But, of course, there are. In newsmaking, there are choices about what to cover, which aspects to cover, where to get video, how to frame it and clip it, what to say, how to smirk, what to say next, and so on. And there are institutional structures that are not really everyday choices – the preference for strong video images, surprising or shocking information, and ongoing institutional biases that shape television news coverage (Bennett, 2006). For this reason, we come to think of something such as the news not as a window on reality, but as a subjective, limited, and constructed depiction that we can or cannot accept as telling us something useful about our social environment. Furthermore, institutions such as mass media news do more than just giving us partial pictures of the world, these enterprises *shape* the world by creating incentives and disincentives for behaviors, favoring and disfavoring people and policies, and, at least partially, defining the public understanding of reality.

And so it goes with other forms of surveillance and information gathering and dissemination. In educational testing, as we have seen, the optics are shaped by budget constraints and the limits of testing technology; they are

shaped by test designer choices about what matters in education; they are shaped by political and technocratic choices about values, metrics, and criteria; and since "testing" is, of course, inextricably linked to the broader structure of American education, they are shaped by the broader histories and patterns of regionalism, racism, classism, and disparity in the American educational system. But beyond the important problems of the optics are the critical effects that such programs have in actually shaping the world they watch. As we have seen, some subjects and types of teaching become disfavored and impracticable in a pedagogical world defined my standardized testing. Testing is not just watching the classroom, it is defining it.

## NOTES

1. For a historical review of the different modes of educational assessment and their transition since the medieval period (see Wilbrink, 1997).
2. But each of these authors, in my view, errs in speaking too confidently about the goals and intentions of NCLB. It is difficult, if not impossible to speak about the "intent" behind legislation that is a hodge-podge of state and federal practices hammered out in compromises between the White House, both houses of the Congress, and a very influential Conference Committee. Add to this the competing agendas created by lobbyists for teachers union and the testing industry, influential members of congress, and education policy entrepreneurs. Next, kick the whole thing out to the 50 states for implementation over a 12-year period with ongoing compliance negotiations and waivers brokered by a federal Department of Education. Then factor in thousands of school districts and schools and their administrators, principals, and teachers. In the end, I would say, it is a mistake to speak confidently about the intent of NCLB as if it were a singularly conceived policy in a hermetically sealed environment.
3. This state autonomy is a fairly bizarre turn, apparently an outgrowth of continued discomfort over a national curriculum and federal government testing regime.
4. As will become clear later, but should be hinted at now, the "standards" approach is just one way to look at educational policy and it is fraught with implications. Under the standards approach, all schools are expected to demonstrably meet the same essential norms of education. Whatever the location, challenges, or makeup of the student body, all children must be proficient at, say, long division by the fourth grade. This means that affluent schools pass the bar with nary a second glance, while schools low in social capital and high social problems may never make the bar. Alternatives, like the value-added approach, will be discussed later.
5. Indeed, many in the educational research community view the state tests as too dubious to use as benchmarks or references because they are so easily gamed by repeat players and the incentives for gaming are so high.

6. The psychometricians are spread even thinner by the use of standardized testing in many professional and military certification programs.

7. This is a joke to rural parents who have just one school per district and to urban parents who live in uniformly poor districts.

8. I can report from personal experience that the High School in my community rewards every sophomore who can pass the NCLB state achievement test the first time through by allowing them to skip all their subject area final exams during the last week of school in the late spring; English, science, math, history, and languages all take this hit in support of the statewide testing. And I can also report what every parent of school-aged children knows – there is not just one test a year, because the schools ready themselves with pre-tests, "short-cycle" assessments tests, and other tests to train and assess the students in preparation for the Big Test.

## REFERENCES

Amrein, A. L., & Berliner, D. C. (2002). The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP test results in states with high school graduation exams. Education Policy Studies Laboratory, Arizona State University. No. EPSL-0211-126-EPRU.

Bennett, L. (2006). *News: The politics of illusion*. New York: Longman.

Foucault, M. (1979). *Discipline and punish*. New York: Vintage.

Gilliom, J. (1994). *Surveillance, privacy, and the law: Employee drug testing and the politics of social control*. Ann Arbor: University of Michigan Press.

Gilliom, J. (2001). *Overseers of the poor: Surveillance, resistance, and the limits of privacy*. Chicago: University of Chicago Press.

Gould, S. J. (1996). *The mismeasure of man*. New York: W.W. Norton.

Haggerty, K. D., & Ericson, R. V. (Eds). (2006). *The new politics of surveillance and visibility*. Toronto: University of Toronto Press.

Hess, F. M. (2003). Refining or retreating? High stakes accountability in the states. In: M. R. West & P. E. Peterson (Eds), *No child left behind? The politics and practice of school accountability* (pp. 23–54). New York: Brookings.

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The unintended consequences of high–stakes testing*. New York: Rowman and Littlefield.

Kohn, A. (2001). Fighting the tests: A practical guide to rescuing our schools. *Phi Delta Kappan*, *82*(5), 348–357.

Lyon, D. (2007). *Surveillance studies: An overview*. New York: Polity.

Monohan, T. (Ed.) (2006). *Surveillance and security: Technological politics and power in everyday life*. New York: Routledge.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston College: National Board on Educational Testing and Public Policy.

Popham, W. J. (2005). *America's "failing" schools*. New York: Routledge.

Rudalevige, A. (2003). No child left behind: Forging a congressional compromise. In: M. R. West & P. E. Peterson (Eds), *No child left behind? The politics and practice of school accountability* (pp. 23–54). New York: Brookings.

Ryan, J. (2004). The perverse incentives of the no child left behind act. *New York University Law Review*, *79*(June), 932–989.
Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed.* New Haven: Yale University Press.
Toch, T. (2006). *Margins of error: The education testing industry in the no child left behind era.* Washington, DC: Education Sector.
Wilbrink, B. (1997). Assessment in historical perspective. *Studies in Educational Evaluation* *23*(1), 31–48.