Slightly revised, September 1, 2012
Comments welcome

School Accountability: To What Ends and With What Effects?

Helen F. Ladd

Edgar Thompson Professor of Public Policy and Professor of Economics
Sanford School of Public Policy
Duke University
Durham, North Carolina 27708
hladd@duke.edu

Keynote address for Conference on "Improving Education through Accountability and Evaluation: Lessons from Around the World," sponsored by the Association for Public Policy Analysis and Management, INVALSI, and the University of Maryland School of Public Policy, Rome, Italy, October 3-5, 2012. The author is grateful to Lucy Sorensen of Duke University for research assistance.

As long as elementary and secondary schooling is primarily publicly funded and mandatory – as it is in all developed countries -- public officials have a responsibility to assure that schools are operating in the public interest. Governments can do so in part through their school funding and curriculum policies, licensure requirements and training programs for teachers, regulations related to the length of the school year or day, and various other regulatory policies. In addition, most, but not all, developed countries also have procedures for holding individual schools accountable for their role in promoting the public interest.

In my talk today, I address two questions. First, **for what** should individual schools be held accountable? Although schools play a central role in the delivery of education to school age children, they are still only one component of a larger education and social system, and that matters for determining for what they should be held accountable.

The second question is: **how well do existing school accountability programs work?** In this section, I focus on two forms of what I refer to as administrative accountability, namely test-based accountability and school inspections. In a test-based accountability system, policy makers hold schools accountable for the performance of their students as measured by test scores or other outcome measures such as graduation rates. In an inspection system, professionally trained inspectors visit schools on a periodic basis with the goal of holding individual schools accountable primarily for the quality of their internal policies and practices. Under either approach unsuccessful schools may well face consequences of various types, including public shaming, additional monitoring, external interventions, or, in the extreme, shutdown of the school. The ultimate goal of an accountability program is to change the behavior of schools, as necessary, in directions consistent with the public interest.

The U.S stands out in relying almost exclusively on test-based accountability. Other countries rely heavily on inspection systems, and some countries have a mixed system, for example, England with its League tables that report student test results and its Office for Standards in Education (OFSTED). The high performing country of Finland is interesting in that it has rejected administrative accountability in favor of reliance on professional accountability among well trained teachers.

Table 1 shows which OECD and other G-20 countries use or do not use school inspections, based on OECD information for 2009. Even though the U.S. does not make heavy use of school inspections, it is included in the OECD data presumably because of the use of school inspections by a few states and cities, most notably New York City. Table 2 provides information on test-based accountability. Determining which countries rely on accountability of this type is challenging because, while many countries have national examinations or assessments, it is not always clear to what extent they use those assessments explicitly for the purposes of school accountability. Table 2 includes the countries that, according to the OECD data, publically report student assessment information by school at the lower secondary level.

I will not be talking today at all about market-based school accountability. Although some people argue that administrative accountability would not be required if education were permitted to operate more like a market – with parents being fully empowered to choose schools for their children and self-governing schools having the power to respond to parental preferences -- that argument is flawed. In a market-driven education system, parents and their children have strong incentives to hold schools accountable for their own private and short term interests but not for the broader public interest. For example, if parents perceive a school to be failing, some will withdraw their children from that school and move them to a stronger school, with little or no attention to the potentially adverse effects of that behavior on the children left behind. So while market based mechanisms may well have a significant role to play in a well-functioning education system, they are not the focus of my talk on public accountability today.

For what should schools be held accountable?

I take quality assurance as the main goal of a school accountability system. That is, the goal is to assure that all schools are as high quality as possible given the resources available to them. The challenge is to make operational the concept of school quality. I suspect that many people would agree that one of the ingredients of school quality is well functioning internal school processes and practices, although they might well disagree on which ones are the most

relevant. More debatable are two alternative potential objects of accountability, namely student outcomes and distributional equity.[1]

In all cases, the guiding principle should be that schools should be held accountable only for things under their control. It would be unfair and would serve no public purpose, for example, to hold schools accountable for inadequate resources if the schools have no control over the level of resources available to them. Moreover, accountability should not be used to push individual schools to achieve standards that are not feasible given the policy tools under their control. Accountability of that type would be unfair to the schools, would not achieve the desired results, and could well induce the schools to behave in counterproductive ways.

Processes and practices.

Historically, individual schools in many countries had very little flexibility to make their own decisions. A tight web of bureaucratic rules and regulations greatly limited the authority that school personnel had over their own operations. Over time, education policy makers have recognized that students have different learning styles and needs, that teaching and learning are complex processes, and that, in the language of economists, there is no well identified production function for education. Acknowledging that what works well in one school environment for certain types of students may not work well in other contexts, policy makers have responded by giving more autonomy to individual schools and in many cases have encouraged, or at least allowed, private groups to establish new types of schools, such as with charter schools in the U.S., foundation schools (formerly grant maintained schools) in England, and Islamic schools in the Netherlands to meet the demands of parents.

It is within this environment of new types of schools and expanded school autonomy that it makes sense for public officials to hold individual schools accountable for their internal processes and practices, with the goal being to assure that the schools are operating in the public interest. As long as the schools are receiving public funding, taxpayers have a right to expect public oversight of how the money is used. Most obvious is oversight of the financial operations of the school to make sure that public funding is not being siphoned off to private purposes. In addition, the public has an interest in assuring that the school is operating in line

[1] For a more complete discussion of measuring school quality, see Ladd and Loeb, 2012.

with national or subnational policies such as curriculum guidelines or teacher qualifications, as well as in line with school specific missions and goals. Further, internal school processes could well include the use of data to understand the needs and progress of individual students, and to make resource allocation decisions designed to promote better student outcomes. Processes of this type are appropriate objects of accountability because they are under the control of school officials.

A further reason to hold schools accountable for internal processes is that some of those processes directly affect children's experiences while they are in school. To the extent that policy makers place a value on the quality of that experience-- which I believe they should, both because of the government's responsibility to children themselves and because their presence in school is required -- any measure of school quality should ideally include some attention to how the students experience the process. That experience depends on how safe the students feel when they are in school, how respectfully they are treated by teachers and other students, how coherently the curriculum progresses from one grade to the next, how well the children are linked to the social services they or their families need, and how engaged they are in their own learning process. Although the specific components of the child's experience may be hard to measure, it should be possible to determine whether a school's practices are consistent with the delivery of a quality schooling experience.

Student outcomes.

Another potential object of school accountability is student outcomes. The main argument for focusing on student outcomes is that internal school processes and practices constitute at best an imperfect and partial measure of school quality. That follows because there is no simple relationship between those policies and student outcomes, and, compared to processes, student outcomes are more directly related to many of the goals of an education system. Moreover, the relationship between processes and outcome undoubtedly differs depending on the school context.

An outcomes based accountability program, however, makes most sense when the student outcomes are broadly defined. If they are narrowly defined, school officials will

undoubtedly shift attention away from other activities in favor of the activities for which they are being held accountable. Such behavior is consistent with a well-known theorem in the literature about the effects of incentive programs in organizations with multiple goals. When only some of the goals can be measured and rewarded, people will focus most of their attention on the rewarded goals to the detriment of the other goals (Milgrom and Roberts, 1992, pp. 228-31; Gibbons 1998). This problem is particularly acute in the field of education because of the multiple benefits of schooling. These benefits include those that accrue both in the short run and the long run to the students themselves in the form of higher achievement, better health and higher earnings, and those that accrue to the more general public in the form of a more educated and informed citizenry.

A major challenge associated with using student outcomes as a proxy of school quality, and hence as an object of accountability, is the attribution problem. Even if current student outcomes that are measurable in the short run, such as test scores or rates of progress through school, are predictive of the long-run schooling outcomes of interest, it is difficult to determine how much of a student's outcomes should be attributed to an individual school. That is true because the benefits of education arise not only from schooling itself but also from families. If one simply uses student test performance as the measure of school quality, one would falsely attribute to schools differences in performance due to natural ability and to family background.

An obvious partial solution would be to hold schools accountable for the gains in student outcomes rather than for outcome levels, so as to avoid holding schools accountable for differences that children bring to school. In fact, however, the evidence shows that ability and family background affect gains as well as levels. In addition, despite the development of sophisticated approaches to estimating the effects of schools on student outcomes, even the most well considered of those does not cleanly isolate school effects from other effects, including for example, those associated with the demographic composition of the school's students.

Importantly, an outcomes-based accountability program is likely to be most productive when, in the absence of accountability policies, schools would be operating inefficiently toward the valued outcomes. To the extent that low student achievement is attributable to other more

systemic factors, such as inadequate funding or impoverished family backgrounds, rather than mismanagement at the school level, outcomes are not an appropriate object of school accountability. Thus, the potential for outcomes-based accountability by itself to improve student outcomes is limited to situations in which educators at the school level are shirking, or are simply not working hard enough or "smart" enough to generate the desired outcomes.

Economists often use the language of the principal agent model to describe this situation. In the context of such a model, the challenge is to set up an appropriate incentive system to induce the agents – in this case, the educators – to operate in ways compatible with the interests of the principal – in this case, state policy makers and the public. The appeal of this model to many policy makers is that it promises gains in student outcomes with little or no increase in resources. Moreover, it does not require that the policy makers understand the education production process. Schools are simply given outcome goals and are charged with figuring out how to meet them.

The problem is that inefficiencies or lack of focus at the school level most likely account for only a small portion of low student performance levels. Instead the low performance may well reflect more systemic problems associated with the operation of a teacher labor market that distributes the weaker teachers to schools serving the most disadvantaged students, inadequate funding for many schools, or, importantly, unaddressed challenges facing disadvantaged children such as poor health or family stress that impede their learning. As a result, placing the burden of raising student achievement on schools alone through an outcomes-based accountability system is likely to generate at most small positive gains in outcomes and could impose large unintended negative side effects.

Moreover, to the extent that the school accountability system sets outcome goals that cannot be achieved through the more effective use of resources alone, it goes against the main guiding principle of school accountability, namely that an accountability system should hold schools accountable only for things within their control. Given the salience of student outcomes, the challenge for policy makers is to design an accountability system is a way that pushes schools in the desired direction without inducing them to engage in counterproductive activities.

Distributional equity.

Yet a third potential object of accountability is what might be called distributional equity, a term that refers to how some groups of children fare relative to others. This object might be interpreted in terms of either outcomes or internal processes and practices. For a number of reasons, distributional equity makes more sense as an object of school accountability when interpreted in terms of how children are treated within the school than in terms of outcomes.

Defined in terms of outcomes, full distributional equity would require equal average outcomes for different groups of students. Given that students from disadvantaged families typically perform at lower levels than those from higher income families, equity would require that schools with greater proportions of disadvantaged students raise the performance of their students by greater amounts than schools serving more advantaged students. In addition it would require that each school aim for equal outcomes for all subgroups of children they serve.

This outcome- based concept of distributional equity – applied either across or within schools -- is not acceptable as an object of accountability given the guiding principle that individual schools should not be held accountable for factors outside their control. At a minimum, the problems of low achievement of some groups would need to be addressed by higher level educational policy makers who have the tools to influence the distribution of students, resources, teachers and programs across schools.

Even the broader set of education policy tools available at a district state or national level are likely to be insufficient, however, to address the barriers to learning faced by many disadvantaged children. No country, even those with very high performing students, has been able to bring the average test scores of its disadvantaged students to the level of its more advantaged students.[2]

That can be seen in Figure 1, which summarizes patterns of tests scores from the Programme for International Student Assessment (PISA) managed by the Organization for Economic Co-Operation and Development (OECD). To facilitate comparisons across developed countries of children from similar backgrounds, the OECD has constructed a measure of the

[2] For further development of this point, see Ladd (2012).

economic, cultural, and social status (ESCS) of the families of all children tested. This measure incorporates information on the household's occupational status, the parents' education level, and, as a proxy for the family's income or wealth, household possessions.[3] This measure is an absolute scale that allows one to compare students with similar family backgrounds across countries.

The figure displays student performance of 15- year olds in reading by ESCS percentile for the U.S and each of the 13 countries whose students scored higher on average than U.S. students in 2009. The reported scores on the vertical axis are standardized as of 2000 to have a mean of 500 and a standard deviation of 100. Note that that the graph tells us nothing about the proportions of children in each category within each country. Instead, it highlights the performance of the typical child at each specified ESCS level in each country

The figure shows strong positive correlations between family ESCS and student performance in all 14 countries. Average test scores for students in the 5th percentile across all the countries are about 350, far below the average of about 660 for students in the 95th percentile, and the test scores rise monotonically both overall and within each country. Even in countries such as Korea, Finland and Canada that are typically viewed as having high performing education systems, the patterns hold: achievement levels of the low ESCS children fall far short of those of their more advantaged counterparts.

Thus, if even high performing countries are unable to close achievement gaps of this type, surely it is not reasonable to expect individual schools to do so. At the same time the variation across countries in the relationship between the achievement between low ESCS and high ESCS children suggests there may be ways to reduce the gap to some extent in many countries.

The issue here is whether there is a role for school accountability to play in narrowing such gaps. My answer is yes because, in the absence of external pressure, schools may well give

[3] The index is based on the following variables: the international socio-economic index of occupational status of the father or mother, whichever is higher; the level of education of the father or mother, whichever is higher, converted into years of schooling; and *an index of home possessions*, which is based on student reports of access to education related possessions such as desks, computers and books, and availability of items such as such as televisions, cars, and cellular phones. The index is standardized to a mean of zero for the population of students in OECD countries, with each country given equal weight. A score of -1.0 on this index means that the student is more disadvantaged than five-sixths of the students in the average OECD country. (OECD, Volume II, 2010, p. 29).

more attention to middle class children than to lower class children simply because middle class parents tend to be more aggressive than their lower class counterparts in looking out for the interests of their children. Given the strong public interest in assuring fair treatment of all students, it would seem appropriate to exert public pressure on individual schools to treat all children fairly with the goal of helping all children reach their potential.

One way to impose such pressure is to hold schools accountable for compiling data on the learning needs and progress of each individual student and for developing and implementing strategies for those who are struggling. One such strategy might well involve working with other community organizations to make sure that children from impoverished families have access to the out of school services, such as physical and mental health care, family counseling services, and enrichment activities they need to perform up to their potential in school.

School accountability in practice

I now turn to the two main approaches to administrative accountability used in practice -- test-based systems that hold schools accountable for measurable student outcomes and inspection systems that hold schools accountable for internal school processes. I begin with the U.S. No Child Left Behind act which provides the clearest example of test-based accountability system and then turn to inspection systems, drawing primarily on my knowledge of the Dutch, New Zealand and English inspection systems (Fiske and Ladd, 2000; Ladd,2010).

Test- based accountability – U.S. style

The quintessential test-based system of school accountability – namely the U.S. No Child Left Behind (NCLB) legislation -- is designed to hold schools accountable both for student outcomes and for distributional equity, with no explicit attention to internal processes and practices. The implicit assumption is that schools will figure out the best ways to change their internal processes to achieve the clear outcome goals specified by the legislation.

Justified in part by the mediocre performance of the U.S. on international test scores, the main goal of the legislation was to increase student achievement. In particular, NCLB, which was implemented in 2002, requires every school to raise the achievement of its students

in math and reading each year to assure that all children reach proficient levels by the 2013-14 school year. NCLB also promotes an equity goal by holding schools accountable not only for the average test scores of all students in the school but also for the test scores of historically underperforming subgroups of students defined by their race or ethnicity, income or disability status. The purpose of doing so is to keep schools from ignoring members of subgroups who might otherwise be left behind.

The longer that individual schools fail to make adequately yearly progress toward the achievement goal specified for all their students and for each of their student subgroups, the more severe are the consequences they face, with the ultimate consequence being school closure. The specific goal of proficiency for all students by 2013/14 was unrealistic from the beginning and continues to be unrealistic. As of 2011, about half of the nation's schools were failing to meet adequate yearly progress (Center for Education Policy, 2011), and the Obama administration has recently been giving waivers to states to relieve them from the requirements of the law.

I argued earlier that placing the burden of raising student outcomes on schools alone is likely to generate at most small positive gains in those outcomes – namely those that are associated with the elimination of teacher shirking and unfocused effort - and could impose large unintended negative side effects. The experience with NCLB bears out these predictions.

In particular, there is little evidence of a change in the trajectory of math and reading test scores on the National Assessment of Educational Progress (NAEP), typically referred to as the Nation's Report Card, after the introduction of NCLB. Using NAEP scores to evaluate the program makes sense both because the sample is nationally representative and because the test itself does not carry high stakes, and is therefore not subject to the same distortions as scores on the high stakes tests used to implement the program in each state. Figure 2 shows eighth grade math and reading scores from 1992 to 2011 with the vertical line in 2002 denoting the beginning of NCLB. Although 8th grade math scores have been rising since 2002, it is hard to attribute that growth to NCLB because it essentially continues the pre-NCLB trend. For reading, the decline in 8th grade scores after 2002 provides no support for a positive effect of NCLB, and even the recent rise in scores has just barely raised them above the pre-NCLB level.

As shown in the next figure, fourth grade test scores present a somewhat more positive picture, especially for math, but still provide no clear evidence of a big impact on test scores. Moreover, even the apparent gain in fourth grade math test scores between 2000 and 2003 occurs too soon to be attributable to NCLB which went into effect for the 2002-2003 school year. Instead, it most likely reflects a variety of other state level policies that were implemented during the 1990s. To be sure some of these state-level policies were test-based accountability programs, but unlike those at the federal level, the state level programs were typically combined with other investments such as additional funding and attention to teacher development and some, such as that in my home state of North Carolina, were better designed in that they were based on gains in test scores rather than the levels in the NCLB legislation.

Of course, the simple NAEP trends by themselves do not tell the whole story because of potentially confounding trends. Several careful and creative studies of NCLB generally confirm the conclusion that NCLB itself has had at most limited positive impacts on student achievement (Wong, Cook and Steiner, 2009; Dee and Jacob, 2011; Cronin et al. 2005). Researchers have had to be creative because of the challenge of evaluating the impact of a single national program that applies to all public schools. They have risen to the task and have used a variety of strategies such as comparing the performance of students in public schools to those in Catholic schools not subject to NCLB, comparing gains in states with low proficiency standards to those in states with high proficiency standards, and comparing the performance of students in states not previously subject to state–based accountability programs with that of students in states newly subject to accountability. Although the results are somewhat mixed, my reading of this more sophisticated analysis is that, consistent with the graphs I just shown, any positive effects on student achievement have been small at best, and, if evident at all, are more apparent for math than for reading.

For example, the national study that generates among the largest effects, that by Dee and Jacob (2011), which is based on data through 2007, finds statistically significant positive effects only for 4th grade math scores, with no positive effects whatsoever for reading and insignificant positive effects for 8th grade math. Moreover, a more recent study by Lee and Reaves that uses a similar comparative interrupted time series methodology but incorporates

more pre-NCLB state-level trend data and more post-NCLB data , finds even smaller effects, with a somewhat different pattern across subjects (Lee and Reaves, 2012).

This finding need not mean that test-based incentives for schools are weak tools for inducing school personnel to change their behavior. Indeed the evidence shows that such incentives can be powerful catalysts for change. The problem is that the changes that schools can make are relatively limited in terms of their potential for increasing student achievement, and some of the induced changes are harmful to the educational process.

The most direct way for individual schools to try to raise math and reading scores is to reallocate resources in favor of those subjects and away from other subjects and other activities, including school recess. As shown in Table 3 which is based on a nationally representative survey of 349 school districts in 2006-07, that is exactly what they have done. The table shows that between 2001 and 2007, schools raised instructional time (measured in minutes per week) in English and math quite significantly, and reduced time for social studies, science, art and music, physical education and recess (McMurrer, 2007; also reported in Rothstein, Jacobsen and Wilder, 2008, Table 3, p. 49).

This shift in resources would not be cause for concern if the true policy goal were to increase math and reading scores alone – although, as measured by performance on the NAEP, it does not seem to have been very effective at achieving even that goal – or if any gains in those subjects were valued more highly than reductions in other subjects. More likely, however, the policy focus on math and reading reflects the practical limitations involved in holding schools accountable for outputs that can be quantitatively measured. In fact, as shown by Rothstein, Jacobson and Wilder (2008), U.S educational stakeholders value a wide range of outcomes that include not just the basic subjects of math and reading but other academic subjects, educational attainment, and development of health behaviors and critical thinking skills that will make them productive in the economic and political life of the country. Hence, the shift in time use is fully consistent with the literature I cited earlier about the dangers of measuring and rewarding only a few of a larger number of valued goals.

A second way for individual schools to raise math and reading scores is to focus on the basic skills that are most readily tested by multiple choice tests rather than on the more

advanced skills necessary for full participation in a changing and knowledge based world. Cross sectional studies of state accountability studies document, for example, a far stronger relationship between the strength of a state's accountability program and performance at the basic level on NAEP than at the proficient level (Carnoy and Loeb, 2002).

Additional evidence of how accountability systems change behavior emerges from studies of how school officials have gamed test-based systems to improve their test results. The most obvious strategy is the selective assignment of low achieving students to special education programs with the goal of removing them from the testing pool, or focusing on students at the margin of attaining a proficient score. Moreover, as various studies have documented with respect to state level accountability programs, schools have also changed their meal programs around test time to increase student achievement and have engaged in selective disciplinary policies toward the same end (see summary in Figlio and Ladd, 2008).

More generally, the test-based incentives under NCLB and other state accountability systems have led to narrow teaching to the test and extensive drilling. Evidence for this point comes in part from the findings that test-based accountability typically is associated with far greater gains in scores on high stakes tests – that is, those that serve as the basis for accountability -- than for low stakes tests. Of even greater concern is the evidence emerging from a number of big U.S. cities and states indicates that some teachers have engaged in outright cheating by altering the test scores of their students[4]. Such behavior, though

[4] Studies have found cheating in response to the high stakes testing under NCLB in New York City; Washington, D.C.; Houston, TX; Atlanta. GA; Pennsylvania among others. *(New York Times, October 17, 2011, http://www.nytimes.com/2011/10/18/nyregion/how-cheating-cases-at-new-york-schools-played-out.html?pagewanted=all); (USA Today, March 30, 2011, http://www.usatoday.com/news/education/2011-03-28-1Aschooltesting28_CV_N.htm); (New York Times, June 11, 2010, http://www.nytimes.com/2010/06/11/education/11cheat.html?pagewanted=all); (New York Times, July 5, 2011, http://www.nytimes.com/2011/07/06/education/06atlanta.html) (New York Times, July 31, 2011, http://www.nytimes.com/2011/08/01/education/01winerip.html?ref=michaelwinerip).*

reprehensible, is not surprising in light of the unrealistic outcome standards imposed by NCLB and the pressures that many schools face to meet the achievement expectations of state and local education officials.

The limited potential for a test-based school accountability program alone to raise student achievement contrasts with the power of a more comprehensive reform effort as illustrated by the performance of the U.S. state of Massachusetts (Table 4). Math NAEP test scores in that state rose by 19 scale points in math for both 4th and 8th graders between 2000 and 2007, for example, -- far exceeding the national gains of 7.2 points in 4th grade and 3.7 points in 8th grade attributable to NCLB by Dee and Jacobs, with similar differentials for reading. A plausible explanation for the larger gains in Massachusetts is that its education reform strategy was more comprehensive than the federal program that focused on test based accountability alone. In particular, the Massachusetts program included a substantial increase in funding (more than doubling in 10 years), new learning standards, revised student assessments based on clear curriculum frameworks, revised teacher licensing and professional development programs, and early childhood programs, as well as parental choice and the creation of new charter schools. Such a package is far closer to what in the education literature has been called standards-based or comprehensive reform than what was implemented under NCLB. Although NCLB is an outgrowth of the standards movement at the national level, it in fact incorporates only one part— the test based accountability part—of what was initially intended to be a far more positive, constructive, and comprehensive approach to raising the achievement of all students.

NCLB also falls short on its second object of accountability, distributional equity. Appealing as subgroup requirements may sound as a component of test- based accountability, they have some serious limitations when applied to individual schools. First and foremost are the statistical problems associated with small numbers. Whenever there are small numbers of students, the chances are high that random variation in student test scores will outweigh any

true signal about gains in test scores (Kane and Staiger, 2002). This concern applies as well to the evaluation of small schools but is particularly serious with respect to subgroups where the numbers are frequently likely to be below 30. For this reason each U.S. state was allowed to set a floor for the minimum size of a subgroup necessary for separate reporting. Thresholds vary from five students in Maryland to as many as 200 in some large schools in Texas. The thresholds provide incentives for districts or schools themselves to keep the size of within-school subgroups below the cut off to the extent they are able, and mean that in practice many schools can still ignore their disadvantaged students. A highly publicized 2006 analysis reported that 1.9 million students were not being counted under their racial and ethnic subgroups because of the thresholds (Bass, Sizon and Feller, 2006). Further as shown by Kane and Staiger (2002), the requirement puts any school with multiple subgroups at a disadvantage relative to other schools and thereby discourages diverse student bodies.

The evidence is at best mixed that the subgroup requirements generate the desired result of higher performance for those subgroups. A careful study of this issue of schools in Texas, which was subject to subgroup requirements well before NCLB found no evidence that holding schools accountable for the performance of subgroups such as African Americans and Latinos had any effect on the performance of those students (Kane and Staiger, 2003). While a more recent study of the subgroup requirement under NCLB in the state of North Carolina, found somewhat more positive results (Lauen and Gaddis, 2012), the recent national study by Lee and Reeves concluded that NCLB had mixed -- and small -- effects on the achievement gaps defined by racial and socioeconomic categories (Lee and Reaves, 2012).

In any case, any positive effects of the subgroup requirements on the achievement of particular subgroups is likely to be quite minor because the requirements do not address in any systematic way the various barriers that impede the learning of disadvantaged students. Without educational policies to assure that schools serving disadvantaged students have access to high quality teachers, and without broader social policies that assure adequate physical and mental health care for children, for example, any external pressure placed on schools to raise the performance of their low performing students is not likely to generate substantial gains in outcomes for such students.

Inspection systems

In contrast to test-based accountability which focuses exclusively on student outcomes, school inspection systems specifically focus primarily on the internal processes and practices within each school, albeit in many cases with some attention to student outcomes. In an inspection system, external review teams visit individual schools on a periodic basis and write public reports highlighting the strengths and weaknesses of the school. The school inspectors are professionally trained and use formal rubrics or protocols designed to produce consistent measures of school quality across units.

In combination with school self-reports and financial and other documents provided to inspectors in advance of their visit, inspectors can gain a relatively solid understanding of the policies and practices within a school. Inspectors typically examine the compliance with rules and regulations, the quality of instruction, student performance and financial management. In addition, in many countries they also look at the satisfaction and perceptions of students, parents and staff.

Direct observation of school quality brings with it its own set of challenges as evident from the well-documented stresses during the early years of the OFSTED program in England (Matthews and Samson, 2004) and the early experience with the new external review program in New Zealand that replaced a far more bureaucratic system of school inspections (Fiske and Ladd, 2000). An innovative part of New Zealand's major reform package of the early 1990s -- a package that turned operating responsibility over to individual schools -- was the establishment of an Education Review Office (ERO) designed to monitor school quality through periodic school visits. The initial intent was to evaluate each school in terms of its own mission statement but the vagueness of the mission statements made that approach unworkable. The ERO instead turned its attention to how well the school was complying with national guidelines for school policies. The reviews were initially heavily focused on management procedures and often became mechanistic, and concerns arose that the standards for success were biased towards the types of procedures that worked best in middle class schools. Starting in 2003, the country introduced a new planning and reporting framework for schools. Under this new system, the ERO now focuses much more on process questions including:

> How well is information on student achievement used, both formally and informally, to develop programs to meet the needs of individuals and groups of students;
>
> How well is available time used for learning purposes; how effective are the systems for identifying and meeting staff professional development needs; and
>
> How well does the school establish partnerships around learning with its community? (Ladd, 2010)

The attention is less on the learning outcomes themselves and more on the robustness and coherence of the internal processes and practices that policy makers believe are associated with good outcomes. Note in addition, the concern with how the school establishes partnerships with other organizations within the community.

This focus on how well schools make use of data on student achievement to allocate resources within schools and on the coherence of policies for supporting student learning emerge as central components of all the inspectorate systems of which I am familiar. The logic underlying such evaluation systems is quite compelling. Because schools differ in terms of the types of students they serve and in their educational strategies, the external reviewers are judging quality based on how well the schools use their resources to address the particular needs of their students. That focus should ultimately promote both higher student outcomes overall, and also the goal of within school equity.

Some inspection systems, such as the one used by the Dutch, also include student test score information in the evaluations. In contrast to the judgements about internal school processes that are made on an absolute scale, however, the Dutch inspectors compare actual test scores with the test scores that would be predicted for the school given the mix of students in the school.

At the same time the inspection approach, like test-based accountability, has drawbacks. One is the danger of placing emphasis on processes that are in fact not directly related to valued educational outcomes, either insofar as the processes have not been validated in the literature as being linked to those outcomes or insofar as the validation is based on measures of quality that are themselves imperfect or narrow. For example, if the

reviewers place a lot of emphasis on the use of data, and those data are all based on tests of basic skills in math and reading, then the observational measure may provide a narrow measure of school quality. Working in the other direction, the review protocol could be so broad that the reports are not useful for distinguishing between those processes that contribute significantly to quality and those that are less directly predictive of quality.

Other limitations are that educators may try to mislead the inspectors about what is really going on in the schools (Wolf and Jannsens, 2007). Further, because it is based on human judgment, the observational approach may be subject to variation in ratings that reflect differences across reviewers rather than differences in education quality. Unless the reviewers are well trained, and the reports tested for reliability across reviewers, the system could provide misleading information on school quality.

The most common concern about the inspection approach, at least in the U.S. discussions, is its high cost. The more frequently schools are visited and the longer are the visits, the higher the costs will be, where costs include the time of the reviewers and the costs of their training as well as the time of school officials for preparing for reviews and responding to them. Presumably it was largely because of the high costs associated with its five-day school visits in the early years of the British program (OFSTED) that the length of the visits were reduced. Similarly the high cost of external reviewers in New York City was one of the reasons that city switched to internal reviewers. At the same time, the fact that New York City, which also has a strong test-based accountability system, also has an inspection system highlights the potential usefulness of holding schools accountable for internal school processes as well as outcomes.

The inspection approach has two clear advantages relative to the test-based approach. The first is that the inspections generate rich information about each school, including information relevant for the experience of children while they are in school. Only with inspections, for example, can policy makers determine anything about the environment within the school that affects the daily life of students and their potential for learning. Not only is such information useful to higher level policy makers, it is also potentially useful to parents as they are making decisions about schools for their children given that parents care about school

environments and understand that some children do better in some types of schools than in others.

The second advantage is that the inspection information is more useful to school managers, and also to any higher level managers, such as superintendents in the U.S. context or managers of groups of schools in the European context. If inspectors deem the quality of a school to be below par on certain dimensions, the school or the higher level managers then have specific information on deficiencies to address. Further, because the information relates to internal school processes and practices, an individual school itself is likely to be a position to change many of those practices. Whether it does address them in fact will depend on the incentives and sanctions built into the overall accountability system, but that is true under either approach.

In general, it is difficult to say whether an inspection system is likely to fare either significantly better or worse than a test-based system with respect to promoting gains in student achievement. Under both approaches the effects are likely to be limited because they are directed at individual schools and therefore do not alter the larger educational context in which the schools are operating. On the one hand, one might expect an inspection system to generate smaller gains in test scores than a test based system because student performance itself, or processes to promote student performance, is only one of many factors that the inspectors examine. On the other hand, the information generated through a well-designed inspection process could be more directly useful to schools, and thereby could potentially reduce within-school inefficiencies more effectively than and without as many adverse effects as with the test-based approach.

Inspection systems differ in the extent to which the inspectors are expected to act as arms-length judges rather than providers of advice. In the latter case, the advice they can provide to an individual school is potentially of high quality given it is based on their observations of effective (or ineffective) practices in many other schools. Even when the inspectors are not permitted to give advice – which is frequently the case so as not to compromise their ability to judge a school – the inspection process can generate useful information for use by the schools in the form of summary public reports on specific topics

based on the information gleaned on successful practices from the many school inspections. Of course, schools may decide to ignore that information unless they face consequences of receiving public reports highlighting their weak internal practices.

Good empirical causal studies of how inspection systems affect student test scores are in short supply. In part this reflects the fact that, unlike the U.S. test based system, the inspection systems in most countries are more focused on quality control than on the goal of raising student test scores, and are often embedded in education systems in which external testing of students on a yearly basis is relatively limited. It also reflects the fact that such studies are hard to do given that all schools are subject to accountability. One recent careful regression discontinuity study of the British OFSTED system by Rebecca Allen and Simon Burgess (2012) reports small positive gains in student test scores two years after schools just failed an inspection visit. The delay in the estimated effects makes sense given the time it takes to change policies. At the same time, the authors point out that the achievement gains emerge only for the schools deemed to have adequate leadership capacity.

A final potential advantage of an inspection system comes from the relationship between the inspection office and the policy makers, a relationship that differs across countries. If the inspection system is run by an independent agency separate from the Department of Education (or Ministry depending on the governance structure in the country), the inspectorate is in a position to write public reports not only about individual schools but also about the quality of the government's policies. In other words, the Inspectorate is able to use its observations in the schools to hold the policy makers publicly accountable for factors that are outside the control of individual schools such as the availability of adequate fiscal resources and adequate training of teaching. That was the situation, for example, in the initial Education Review Office in New Zealand. Even in systems in which the Inspectorate has less autonomy, as in the Netherlands, it can still productively contribute to the policy making process through its access to comprehensive information on what is happening at the school level.

Conclusion.

So what can we conclude about the two types of school based accountability systems?

One conclusion is that there are some serious downsides to a badly designed test based accountability system, which is the way I would characterize the 10- year experiment with No Child Left Behind in the United States. Any school accountability system that labels half of all U.S. schools as failing, that generates very little in the way of achievement gains even in the subjects that are the focus of the accountability effort, and induces large negative unintended side effects is not to be recommended to other countries.

A second conclusion is that neither approach to school accountability -- the inspection approach nor the test based approach – by itself is likely to generate significant increases in the achievement of students, at least as measured by student performance on low-stakes tests rather than the high-stakes tests used in the accountability system itself. Although school-based accountability still has a role play as part of a larger strategy for increasing student achievement, that role is primarily to assure that all schools meet quality standards, and are operating efficiently toward the desired achievement and other goals of a well-designed education system. Stated differently, accountability pressure on individual schools is no substitute for broader and more systemic policies and investments designed to raise student achievement.

Third, despite their cost, school inspections should be part of any effort to hold schools accountable. This conclusion follows in part because the quality of the school environment itself is one of the valued benefits of schools and one for which the government has a particular responsibility to protect. The only way to know what is really going on in schools is to visit them. In addition, the conclusion follows because the inspection process, including the public inspection reports, provides potentially far more useful information for schools and other stakeholder than does the test-score information available from a test based system.

Finally, there is no single best model of an inspection system. All the systems of which I am aware are continually making changes and evolving in new directions. I, personally, would like to see far more experimentation with the inspection model in the United States with different models tried in different states. One advantage of that such experimentation is that it would provide useful information for researchers such as me, and many of you at this conference, to do the research necessary to learn what strategies work best in what contexts.

References.

Allen, R. and S. Burgess. 2012. "How should we treat underperforming schools? A regression discontinuity Analysis of School Inspections in England." DoQSS Working Paper no. 12:02. Institute of Education, University of London.

Bass, F., N. Dizon, and B. Feller. 2006. "Schools skirt "No Child Left Behind" rule. Associated Press, April 17.

Carnoy, M. and S. Loeb. 2002. "Does external accountability affect student outcomes? A cross-state analysis." *Education Evaluation and Policy Analysis*. Vol. 24(4), pp. 305-331.

Cronin J., G,G. Kingsbury, M.S. McCall and B. Bowe. 2005. *The impact of the No Child Left Bhind Act on student achievement and growth, 2005 edition*. Northwest Evaluation Association, Northwest Evaluation Association Technical Report.

Dee, Thomas and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Outcomes," *Journal of Policy Analysis and Management*, Vol 30, no. 3, pp. 418-446.

De Wolf, I.F. and F.J.G. Jannsens. 2007. "Effects of Inspection and Accountability in Education." *Oxford Review of Education*. Vol.33(3), pp. 379-396.

Figlio, D. and H. Ladd. 2008. "School accountability and student achievement." In H. Ladd and E.Fiske (eds), *Handbook of Research in Education Finance and Policy.* New York and London: Routledge.

Fiske, Edward B. and Helen F. Ladd. 2000. When Schools Compete: A Cautionary Tale. Washington, D.C.: Brookings Institution.

Gibbons,R. 1998. "Incentives in Organizations," *Journal of Economic Perspectives, 12(4), pp. 115-132.*

Kane, Thomas and Douglas Staiger. 2002. "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." *Brookings Papers on Education Policy 2002.* Washington, D.C.: Brookings. PP. 235-83.

Kane, Thomas and Douglas Staiger. 2003. "Unintended Consequences of Racial Subgroup Rules." In Paul E. Peterson and Martin West, eds, *No Child Left Behind? The Politics and Practice of School Accountabiity.* Washington, D.C. : Brookings. Pp. 152-176.

Ladd, Helen F. 2010. "Education Inspectorate Systems in New Zealand and the Netherlands," *Education Finance and Policy*, vol. 5, no. 3.

Ladd, Helen F. 2012. "Education and Poverty: Confronting the Evidence." (Presidential address to the Association for Public Policy and Management, 2011) *Journal of Policy Analysis and Management,* Spring.

Ladd, H. and S. Loeb. 2012 (in press). "The Challenges of Measuring School Quality: Implications for Educational Equity." In D. Allen and R. Reich (eds.), *Education, Democracy, and Justice*. University of Chicago Press.

Lauen, D. and M. Gaddis. 2012. "Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability on student achievement." Vol. 34(2), pp. 188-208/

Lee, Jaekyung and Todd Reaves. 2012. "Revisiting the Impact of NCLB High-Stakes School Accountability, Capacity and Resources: State NAEP 1990-2009 Reading Achievement Gaps and Trends." *Education Evaluation and Policy Analysis.* Vol. 34 (2), pp. 209-231.

Matthews, Peter and Pam Sammons. 2004. *Improvement through inspection: An Evaluation of the impact of Ofsted's work*. Institute of Education, University of London, (July).

McMurrer, Jennifer. 2007. Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era*. Washington, D.C.: Center on Education Policy, July (revised December).*

Milgrom P. and J. Roberts. 1992. *Economics, Organization and Management.* Prentice Hall.

Rothsetin R., R. Jacobsen, and T. Wilder. 2008. *Grading Education: Getting Accountability Right.* Washington, D.C.: Economic Policy Institute.

Wong, M. ,T.D. Cook, and P.M. Steiner. 2009. "No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time serviced each with its own non-equivalent comparions series. Northwestern University Institute for Policy Research, Northwestern University. Working Paper Series WP-09-11.

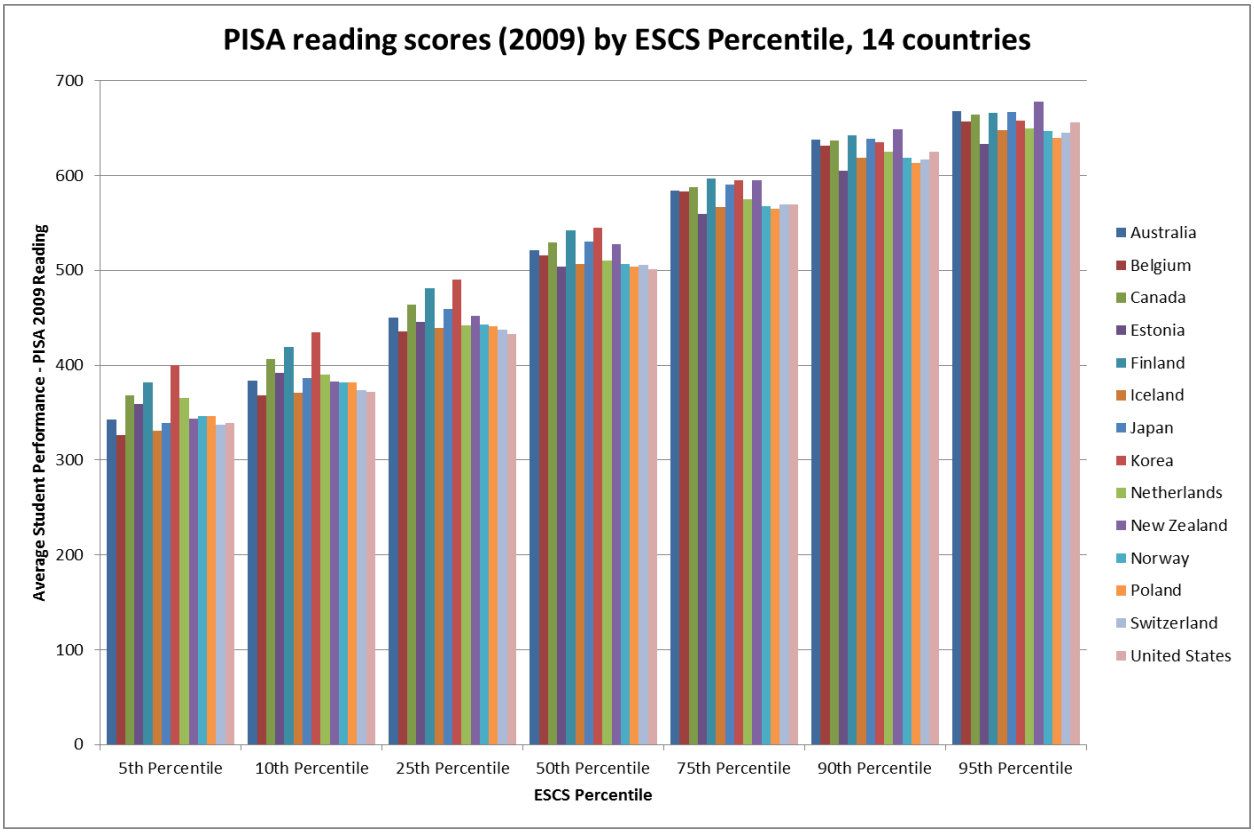
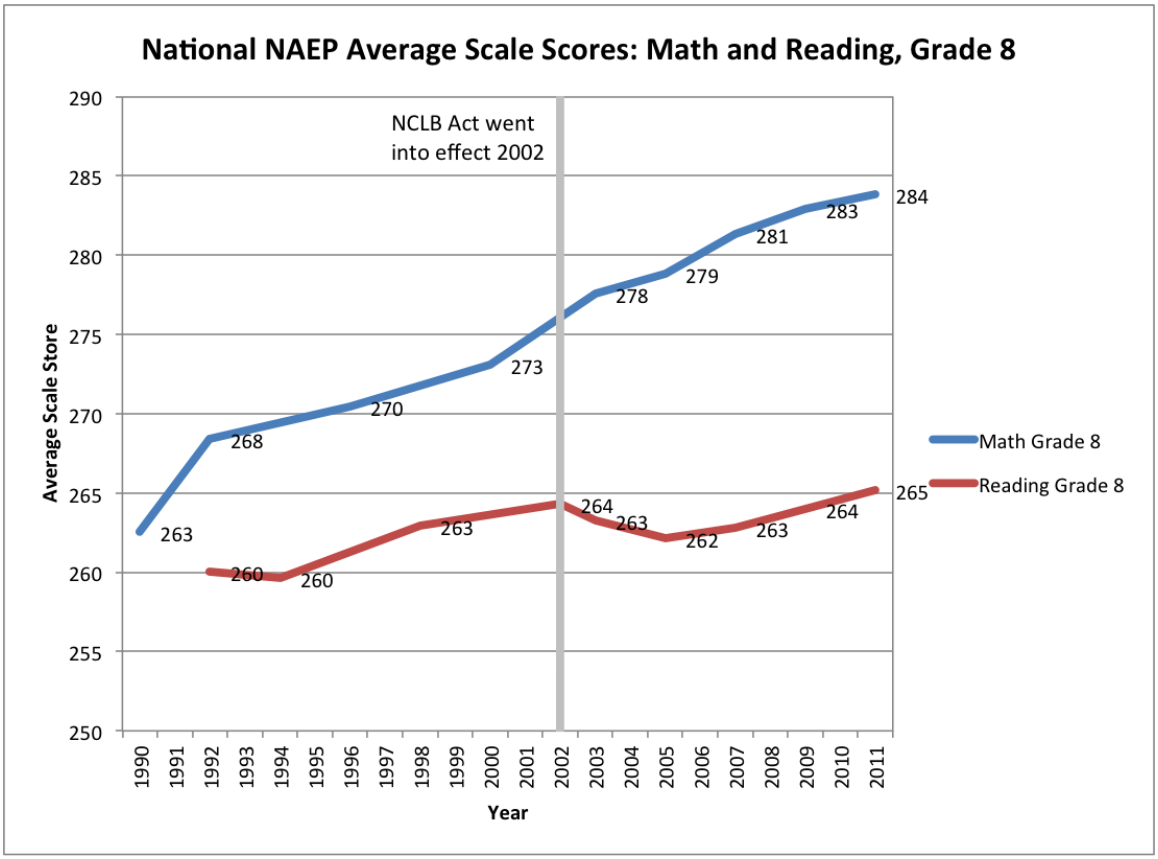

PISA reading scores (2009) by ESCS Percentile, 14 countries
Average Student Performance - PISA 2009 Reading
700
600
500
400
300
200
100
0
5th Percentile
10th Percentile
25th Percentile
50th Percentile
75th Percentile
90th Percentile
95th Percentile
ESCS Percentile
Australia
Belgium
Canada
Estonia
Finland
Iceland
Japan
Korea
Netherlands
New Zealand
Norway
Poland
Switzerland
United States

Figure 1.


National NAEP Average Scale Scores: Math and Reading, Grade 8
NCLB Act went into effect 2002
Average Scale Store
Year
Math Grade 8
Reading Grade 8
290
285
280
275
270
265
260
255
250
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
263
268
270
273
278
279
281
283
284
260
260
263
264
263
262
263
264
265
Figure 2.

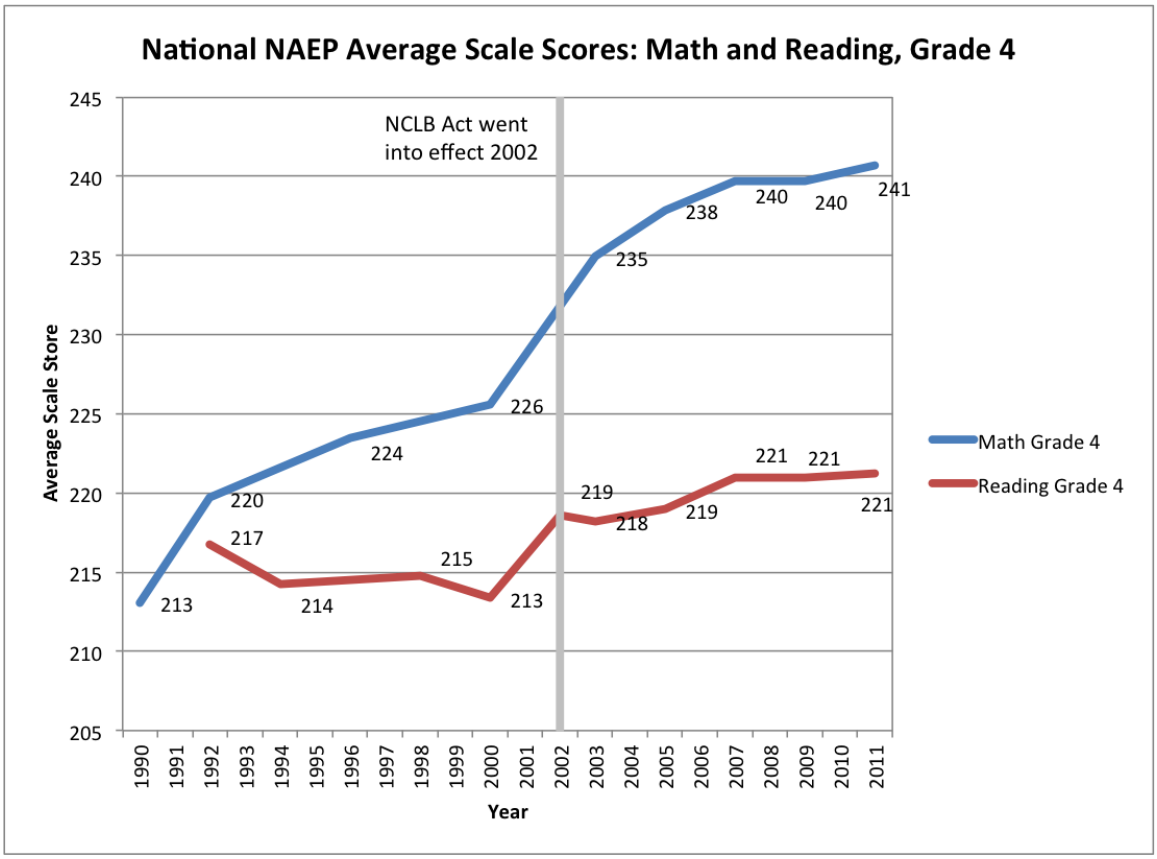

National NAEP Average Scale Scores: Math and Reading, Grade 4
NCLB Act went into effect 2002
Average Scale Store
Year
Math Grade 4
Reading Grade 4
245
240
235
230
225
220
215
210
205
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
213
220
224
226
235
238
240
240
241
217
214
215
213
219
218
219
221
221
221

Figure 3.

Table 1. Countries with and without school inspections (OECD and other G20 countries), lower secondary schools, 2009.	
Countries with school inspections	**Countries without school inspections**
Austria	Denmark
Belgium (Fl)	Finland
Belgium (Fr)	Greece
Czech Republic	Hungary
England	Italy
Estonia	Japan
France	Mexico
Germany	
Iceland	
Ireland	
Israel	
Korea	
Netherlands	
New Zealand *	
Norway	
Poland	
Portugal	
Scotland	
Slovak Republic	
Spain	
Sweden	
U.S. **	
Source . OECD (2011) Education at a Glance, 2011: OECD Indicators, Chart D5.4a. *No information provided to OECD, but included in list because they have school inspections. ** Limited to a few states and cities. Countries with no information or for which the category does not apply: Austria, Canada, Slovenia, Switzerland, Luxembourg, Turkey.	

Table 2. Countries that report level of school performance for most recent year on national assessments, lower secondary level, 2009. (Based on OECD and other G-20 countries that provided information).

Australia	Mexico
Belgium (FL)	Norway
Chile	Slovak Republic
Germany	Spain
Hungary	Sweden
Iceland	United States

Source. OECD, (2011) Education at a Glance, 2011: OECD Indicators, Chart D5.2a

Table 3. Changes in instructional time, districts with at least one NCLB-sanctioned school	
Subject area	**Average weekly change in instructional time (minutes), 2001-2007**
English	**+163**
Math	**+86**
Social Studies	**-90**
Science	**-94**
Art and music	**-61**
Physical education	**-57**
Recess	**-60**
Source. McMurrer, 2007, Table 4. Also reprinted in Rothstein, Jacobsen, and Wilder, 2008.	

Table 4: Gains in NAEP Scale Scores in Massachusetts compared to National Estimates of Effects of NCLB

	Math	
	Fourth grade	**Eighth grade**
Massachusetts (2000-2007)	**+19.0**	**+19.0**
NCLB effect (Dee and Jacobs)	**+ 7.2**	**+ 3.7 (NS)**
	Reading	
Massachusettts (2000-2007)	**+13.0**	**+ 4.0**
NCLB effect (Dee and Jacob)	**+ 2.3 (NS)**	**-2.1 (NS)**

Sources: National Center for Education Statistics; Dee and Jacobs, 2011.
NS signifies not statistically significant. The 2000 scale scores for Massachusetts were in math 233 (fourth grade) and 279 (eighth grade) and in reading 223 (fourth grade) and 269 (eighth grade.