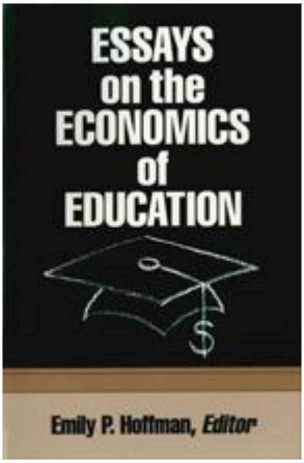
W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH

Upjohn Institute Press

Can Schools Be Held Accountable for Good Performance?

A Critique of Common Educational Performance Indicators



Robert H. Meyer
University of Chicago

Chapter 4 (pp. 75-110) in:
Essays on the Economics of Education
Emily P. Hoffman, ed.
Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 1993
DOI: 10.17848/9780880995542.ch4

Can Schools Be Held Accountable for Good Performance?

A Critique of Common Educational Performance Indicators*

Robert H. Meyer
University of Chicago

Educational indicators are increasingly being used to assess the efficacy of American education. Local newspapers regularly report how students in local schools perform on nationally standardized tests, and a growing number of states publish formal school report cards that provide an assortment of student outcome, enrollment, and financial indicators. In April 1991, President Bush elevated the discussion of educational indicators to the national level with "America 2000," a proposal to establish a national examination system, complete with school district, state, and national report cards (U.S. Department of Education 1991).

The growing demand for educational performance indicators has been motivated in large part by a growing demand for public *accountability* defined in terms of hard outcomes, such as standardized test scores, rather than inputs, such as teacher qualifications, class size, and course requirements. Demands for public accountability have been particularly strong in states that have dramatically increased expenditures on education and in states that have launched major school improvement efforts. The increased demand for public accountability in elementary and secondary education has paralleled similar demands for increased accountability in other public sector activities, for example, the Job Training Partnership Act and the new JOBS program, enacted as part of the Family Support Act.

*This research has been supported by the La Follette Institute of Public Affairs and the Institute for Research on Poverty, University of Wisconsin-Madison. The author has benefited enormously from discussions with Bill Clune, Jeff Dommitz, and Andy Porter and from comments by Ken Bickers, Mike Wiseman, and participants in the LaFollette Institute public affairs workshop.

Despite the groundswell of interest in data on school performance, many educators and scholars fear that poorly implemented performance indicators could ultimately be worse than no indicators at all. These fears are not groundless. As will be discussed in this paper, performance indicators based on achievement tests could be flawed in two major ways. First, the achievement test underlying a performance indicator could be susceptible to "narrow" teaching to the test or could fail to reflect a school's true educational objectives. Second, a performance indicator constructed from a simplistic or otherwise inappropriate statistical model could fail to reflect the true contribution of a school to growth in measured student achievement. Under these conditions, a high stakes system of educational performance indicators could severely distort the behavior of educators and students.

The purpose of this paper is to assess the statistical adequacy of the most commonly used educational performance indicators. One of the major conclusions of the analysis is that the typical indicators used to assess school performance—average and median test scores—are highly flawed as measures of school performance. As a result, they are of limited value, if not useless, for evaluating relative school performance or school performance over time. Indeed, simulation results indicate that changes over time in average test scores could very well be negatively correlated with actual changes in school performance.

The analysis also demonstrates that the typical indicators used to assess school performance are likely to provide schools with the perverse incentive to "cream," that is, to raise measured school performance by educating only those students who tend to have high test scores. The potential for creaming is apt to be particularly strong in environments characterized by selective admissions. However, creaming could also exist in more subtle, but no less harmful, forms. For example, schools could create an environment that is relatively unsupportive for potential dropouts, academically disadvantaged students, and special education students, thereby encouraging these students to drop out or transfer to another school. Alternatively, high-quality teachers and administrators could gravitate to neighborhood schools that predominantly serve high-scoring students.

The paper is.organized in nine major sections, the first of which is this introduction. The second section is a discussion of the problems that exist with traditional standardized tests; the third presents an assessment of the validity of the average test score. I demonstrate that this commonly used indicator is highly flawed as an indicator of school performance. In the fourth section, I demonstrate that an alternative indicator, the gain indicator, avoids all but one of the major flaws associated with the average test score. In particular, the gain indicator fails to measure the value-added contributions of schools to growth in academic achievement. The seventh and eighth sections draw on simulated and actual data to illustrate the advantages of gain indicators over average summary scores. I first investigate value-added indicators, and then consider the consequences of evaluating schools on the basis of incomplete indicators. Finally, I present recommendations for the phased-in development of valid educational performance indicators. An appendix provides technical information concerning the simulations reported in the fifth section.

The Problems With Traditional Standardized Tests

Many educators and testing experts believe that there is a great need for new and improved ways of testing student achievement. A major problem with national standardized tests is that they are designed to appeal to all schools regardless of their educational objectives. These tests, if used in a high stakes indicator system, could drive teachers and administrators to focus almost exclusively on low-level academic content (Smith and O'Day 1990; Clune 1991). The achievement tests used as the basis for a performance indicator system should ideally reflect a balance of low- and high-end content so that the performance of schools that serve low- and/or high-achieving students can adequately be measured. This implies that a minimum competency test is unlikely to be satisfactory as the basis for measuring school performance. The problem with minimum competency exams is that many students receive a perfect score year after year. If the tests differ from one grade to the next, the recorded gain for these students is totally artificial. If the tests do not differ, their recorded gain is zero—in most cases, a vast understatement of their true gain in achievement. The simple achievement models presented later in this paper are not really appropriate for tests that exhibit low ceilings and/or high floors. However, the models could be extended to allow for the "censoring" of test scores at the high and low extremes of the test score distribution.

Critics of standardized tests also argue that conventional multiple choice tests are not well suited to assessing skills involving higher order thinking and problem solving, the kinds of skills that are increasingly valued in our economy. They argue that the multiple choice format is generally limited to asking simple questions that have definite answers. As a result, a history exam is reduced to questions about dates and events, rather than the causes of the Civil War; a mathematics exam is reduced to a long series of addition and multiplication problems, rather than questions involving the application of mathematics to solving real-world problems. It is feared that a system of performance indicators based on such tests is likely to encourage teachers and administrators to focus their teaching on repetitive, rote learning.

These criticisms have stimulated a number of states to begin developing new, performance-based tests (Dominitz and Meyer 1991). One commonplace example of an authentic performance-based test is the field portion of a driving test. A driving test assesses, more or less, what a driver needs to know to drive on city streets. Indeed, the best way to pass a driving test is to practice driving. In contrast, typical standardized math tests fail to assess what most students need to know about mathematics, the capacity to tackle extended real-world problems calling for the application of diverse mathematics skills. Advocates of performance-based tests argue that these tests will be relatively immune to the phenomenon of narrow teaching to the test and more congruent with state educational curriculum goals.

Level Indicators

Standardized student testing is conducted for a variety of different reasons: to provide information on individual students and obtain aggregate school-level indicators. At the student level, for example, standardized test scores may be used to diagnose student strengths and weaknesses in subskill areas,[1] to guide teachers in providing instruction that matches the needs of individual students, to guide students in making curriculum and career choices, to determine, in states that have minimum competency examinations, whether students are eligible for graduation, and to guide postsecondary institutions and employers in making admissions and hiring decisions, respectively.

These data, if aggregated to the classroom or school level, yield educational indicators that measure, for example, the share of students scoring above or below certain thresholds or the average level of achievement. I refer generally to statistics of this kind as level indicators. As previously mentioned, level indicators are widely reported by schools. Indeed, they are calculated and readily made available by the companies that provide testing services to schools throughout the nation (Goldman 1990). They are also reported at the national level by the National Assessment of Educational Progress. Unfortunately, some of the level indicators reported by schools and states are subject to obvious statistical flaws. Well-known examples include average SAT and ACT scores. The problem with these indicators is that they are based on nonrandomly selected groups of students—in particular, those students who aspire to attend selective colleges or universities. As discussed by Hanushek and Taylor (1990), Powell and Steelman (1984), and Wainer (1986), these indicators tend to be highly unreliable as measures of the true level of achievement in schools and states. In this paper, I limit my analysis to level indicators that are not subject to these problems.

If correctly constructed and based on appropriate tests, level indicators convey potentially useful *descriptive* information concerning the proficiency levels of students in particular classrooms or schools. Indeed, they could sensibly be used to target assistance (financial or otherwise) to schools that serve students with low test scores. The critical question for the present discussion is whether such indicators are valid and useful measures of school or classroom performance. The answer to this question is no. School performance indicators, by definition, must validly measure the contribution of schools to growth in student achievement for students in particular grades or sequence of grades.

Average (or median test) scores fail to do this for four reasons. First, the average test score fails to *localize* school performance to a specific classroom or grade level—the natural unit of accountability in a traditional school.[2] This lack of localization is, of course, most severe at the highest grade levels. In my judgment, a performance indicator that fails to localize school performance to a specific grade level or classroom is likely to be a relatively weak instrument of public accountability.

Second, the average test score reflects information that is *aggregated across time and grade levels* and therefore tends to be grossly *out of date*. For example, consider the average test score for a group of high school seniors. The test scores for these students reflect learning that occurred in kindergarten, roughly twelve-and-one-half years earlier, through the twelfth grade. Indeed, a twelfth-grade level indicator could be dominated by information that is ten or more years old.[3] The fact that average test scores reflect out-of-date information severely weakens them as instruments of public accountability. In order to allow educators to react in a timely and responsible fashion, performance indicators must reflect information that is current.

Third, average test scores at the school, district, and state levels tend to be highly *contaminated* due to student mobility in and out of different school systems. For example, the typical twelfth-grade student is likely to attend several different schools over the period spanning kindergarten through twelfth grade. For this student, a test score reflects the contributions of more than one and possibly many different schools. The problem of contamination is compounded by the fact that rates of student mobility tend to differ dramatically across schools. Contamination is apt to be especially high in communities that undergo rapid population growth or decline or experience significant changes in their occupational and industrial structure. Contamination due to student mobility is probably a relatively minor problem at the national level, since rates of in- and out-migration are low compared to rates of mobility within the nation.

Fourth, the average test score is not a *value-added* indicator; that is, it fails to measure the distinct contribution of a school to growth in educational achievement. As a result it absorbs differences across schools in student achievement levels that are due not to differences in school productivity but rather to variations in student achievement prior to entering school and to differences in growth in student achievement that are systematically related to differences in student and family background characteristics.

In summary, the average test score suffers from four major flaws, any one of which could be sufficient to invalidate it as a measure of school performance. In the next section I therefore consider an alternative indicator that largely avoids the problems of nonlocalization, aggregation across time and grade levels, and contamination, namely, the gain indicator. Immediately following is a series of simulations that compare the average test score relative to the gain indicator.

Gain Indicators

The gain indicator measures the average growth (or gain) in achievement from one point in time to another for a given cohort of students. If students are tested at least once a year, the gain indicator largely avoids three of the problems that seriously undermine the average test score as a valid and up-to-date measure of school performance: the problems of nonlocalization, aggregation across time and grade levels, and contamination due to student mobility. However, the gain indicator does not measure the value-added contribution of schools to growth in student achievement, that is, it does not measure school performance. Rather, it measures the joint contributions of students, families, communities, and schools to growth in student achievement. As such, it is an extremely informative *descriptive* indicator that should be included, along with the value-added indicators introduced below, in a comprehensive system of educational indicators.

The quality of the gain indicator depends critically on the frequency of student testing. Annual (or more frequent) testing is ideal for several reasons. First, performance is localized to single grade levels, the natural unit of accountability. Second, the information reflected in the indicator is completely up to date. Third, contamination due to student mobility is limited only to students who transfer schools during the school year.

As the time interval between tests increases, the problems of localization, contamination, and aggregation over time and grade levels become more acute. In fact, for time intervals of more than two years, it could prove difficult to construct valid and reliable gain indicators for schools with high mobility rates. There are two options in such cases. First, mobile students could simply be excluded from the data for a classroom or school. Dyer, Linn, and Patton (1969) refer to this as the "matched sample" approach. The problem with this approach is that nonmobile students are apt to be unrepresentative of the school population as a whole, both in terms of student characteristics and educational experiences. Moreover, the number of nonmobile students in such cases could be simply too small to yield reliable (statistically precise) estimates of average student gain. The second option is to include mobile students in the gain comparison for a given school even though the students obtained part of their schooling from another school. Of course, this option is feasible only if mobile students take the same tests in different schools and if their test scores are made available to the schools to which they move or exit. This clearly would be feasible only in states that have mandated state assessment systems. Even so, students who move across state lines would be lost unless the states happen to use the same state tests and are prepared to exchange student test data. A more fundamental problem with this approach is that the contamination introduced by mobile students severely jeopardizes the validity of the gain indicator if the mobility rate is high. The bottom line is that infrequent testing seriously compromises the validity of the gain indicator.

How Bad is the Average Test Score as a Measure of School Performance? Simulation Results

This section presents a series of simulations designed to assess whether the average test score has *any* value as a measure of educational productivity. I consider the validity of the average test score for comparisons across schools and for comparisons over time for the same school. The second type of comparison is particularly relevant for the purposes of evaluating the efficacy of school reform efforts.

Let $L(c, g)$ represent the average level of achievement in a particular school for cohort c at the end of grade g. Similarly, let $G(c, g)$ represent the average gain in achievement in a particular school for cohort c from the end of grade $(g\text{-}1)$ to grade g, that is,

$$G(\mathrm{c}, g) = L(c, g) - L(c, g\text{-}1). \tag{1}$$

Equation (1) implies that, for a given cohort, the average level of achievement at the end of a particular grade, say grade 10, is the sum of prior gains in achievement plus the initial average level of achievement, that is,

$$L(c, 10) = L(c, 0) + G(c, 1) + \ldots + G(c, 10). \tag{2}$$

Given alternative assumptions concerning initial achievement and the pattern of gain values over time and across grade levels, I can compute the average level of achievement at the end of grade 10 for each cohort.

To emphasize the contrast between average achievement and the gain in achievement, I assume that average initial achievement and average student characteristics are identical for all schools at all points in time.[4] I also assume, for simplicity, that all students begin first grade at the same age and advance from one grade to the next each year. In this case, a unique time index is implied by the cohort and grade level indices. The relationship between time, birth cohort, and grade level is given by the formula[5]

$$t = c + g + 6.$$

To facilitate comparisons across schools at the same grade level, I standardize the school gain values so that the average gain for the entire population at a given point in time is equal to zero at each grade level. Average 10th grade achievement is similarly standardized to have mean zero. Finally, I assume that the achievement test underlying this analysis is scaled so that the standard deviation of school gain values is approximately equal to 10 in the typical grade.[6] To provide the reader with some intuitive sense of the standardized gain values, table 1 lists percentile values associated with a range of gain indicator values.

The first pair of simulations illustrate the failure of average test scores to localize school performance to specific grade levels. Subsequent simulations illustrate the consequences of contamination and aggregation across time and grades. Technical details of the simulations are presented in the appendix.

Table 1. Gain Percentile Values, Given the Assumption that Average Gains are Normally Distributed

Gain indicator values, given zero mean and standard deviation equal to 10	Gain percentile values, given the assumption that average gains are normally distributed
30	99.9
20	97.9
10	84.1
5	69.2
0	50.0
-5	30.8
-10	15.9
-20	2.3
-30	0.1

The first simulation, as summarized in table 2, contrasts three schools that differ in terms of their patterns of (standardized) gain in grades one through six and grades seven through ten, respectively. To simplify the analysis I assume that these patterns persist over time and that there is no student mobility. School 1 exhibits gain values of zero (the average) at all grade levels. School 2 exhibits exceptionally high gain values in the upper grades and exceptionally low gain values in the lower grades. Finally, school 3 exhibits a pattern of gain values that is exactly opposite to the pattern exhibited for school 2. As indicated, the three schools differ fundamentally in terms of their gain values in the early and late grades. Despite these differences, however, the schools are indistinguishable in terms of their average level of achievement at the end of tenth grade. The exceptionally high and the exceptionally low gain values simply cancel out for schools 2 and 3.

A similar result is observed in the second simulation, as depicted in figure 1. Figure 1 charts the average level of tenth-grade achievement over time, prior to and after the implementation of hypothetical academic reforms in 1992. The academic reforms are assumed to follow an era of stable but average gains in achievement at all grade levels. Panels A and B in figure 1 depict two different scenarios. In panel A the average achievement gains at each grade level increase gradually after 1991. In panel B, the average achievement gains also increase steadily, but they are limited to grades seven to ten. The gain values are indicated on the graph by the solid gray lines. The tenth-grade achievement levels are indicated on the graph by the solid black lines. As in the previous simulation, the two schools differ substantially in terms of their gain values at different grade levels. Despite these differences, however, there is no perceptible difference between the two schools in terms of average tenth-grade achievement. In short, these two simulations demonstrate that average test scores provide no information on differences in productivity between different levels of a school system. They do, however, suggest that average test scores provide at least a rough indication of the productivity of the school system overall. In fact, this is generally not true, as is demonstrated below.

Table 2. Average Tenth-Grade Achievement by School, Given Alternative Patterns of Gain

		Average gain per grade		
School	Initial achievement	Grades 1 to 5	Grades 6 to 10	Average achievement at the end of tenth grade
1	0	0	0	0
2	0	20	-20	0
3	0	-20	20	0

The second set of simulations illustrates the problem of aggregation across time and grade levels. These simulations demonstrate vividly how average test scores are determined in large part by past gains in achievement and hence are apt to be quite misleading as indicators of current gains. To highlight the problem of aggregation across time and grade levels I assume that achievement gains within a school are identical at all grade levels and that there is no student mobility. Figure 2 charts average tenth-grade achievement and average achievement gains over time, prior to and after the introduction of hypothetical academic reforms in 1992. Panel A of figure 2 depicts a scenario in which academic reforms reverse a trend of gradual deterioration in average achievement gains across all grades and initiate a trend of gradual improvement in average achievement gains across all grades. Panel B

Figure 1. Average Tenth-Grade Achievement Given Alternative Patterns of Gain in Grades 1 to 6 and 7 to 10

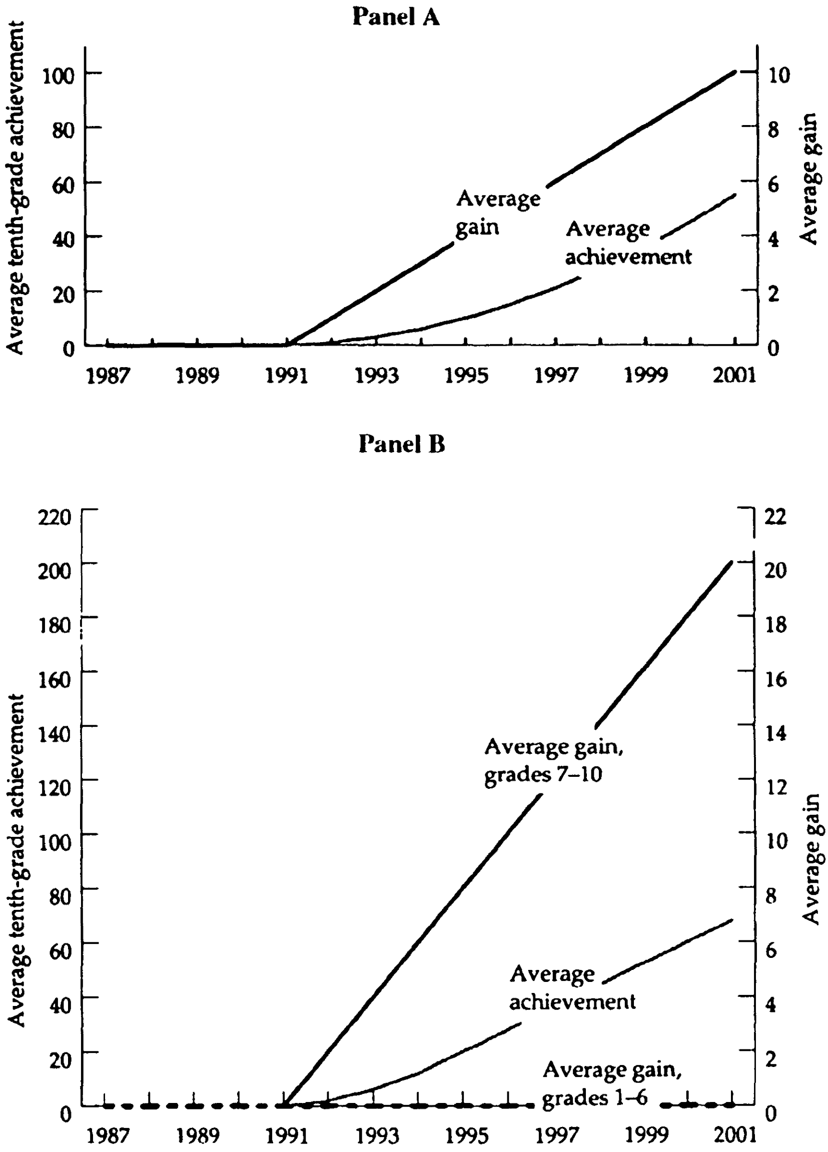
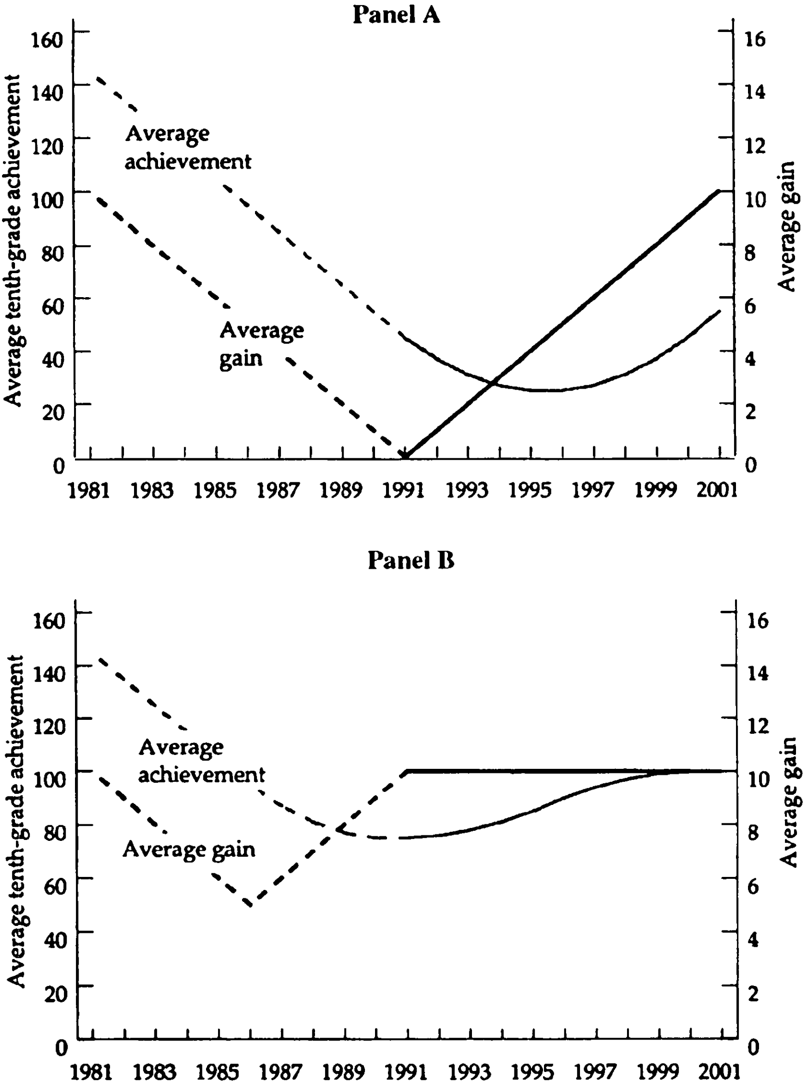


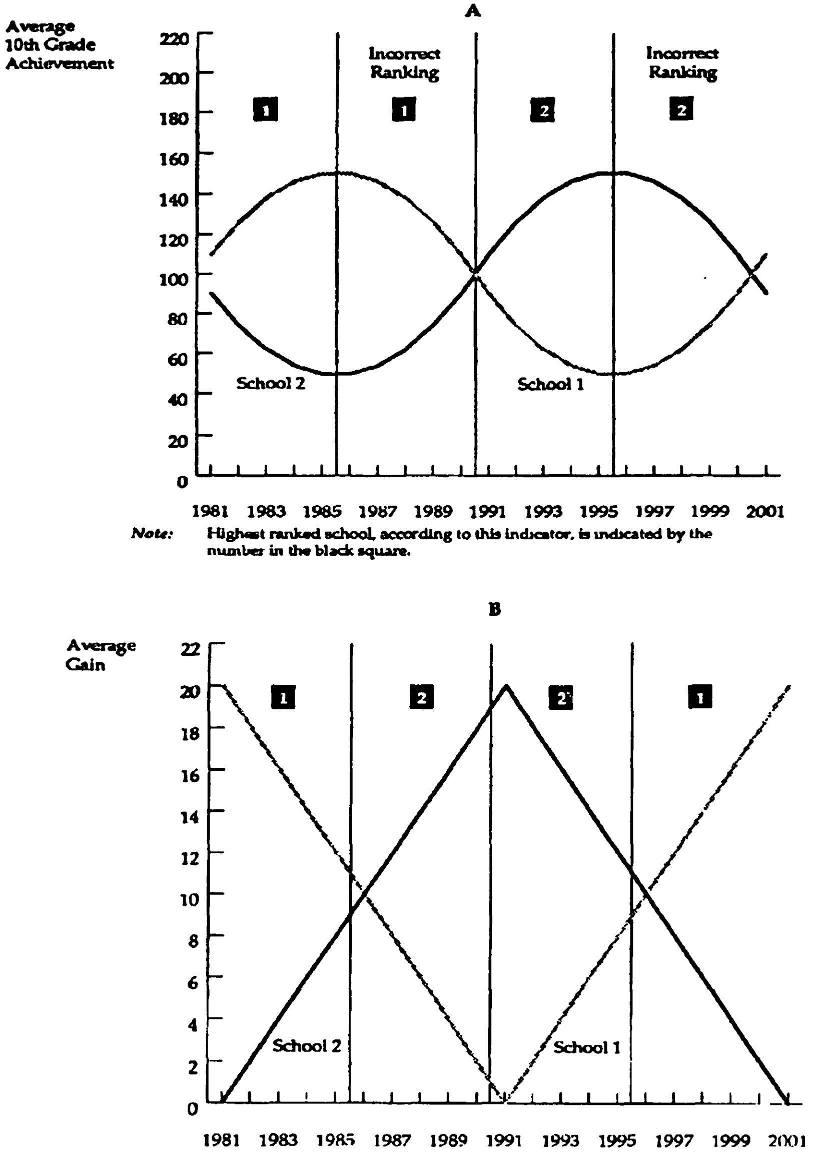
Figure 2. Average Tenth-Grade Achievement Given Alternative Patterns of Gain Over Time



of Figure 2 depicts a scenario in which academic reforms have absolutely no effect on average achievement gains. The reforms, however, are preceded by an era of gradual deterioration in average achievement gains across all grades, followed by a brief period (1987–1991) of gradual improvement across all grades. As indicated in the graph, the average tenth-grade test score provides a totally misleading view of the effectiveness of the hypothetical academic reforms implemented in 1992. In panel A, the average 10th grade test score *declines* for five years after the introduction of successful reforms. In panel B, the average tenth-grade test score *increases* for a decade after the introduction of reforms that have no effect on student achievement growth. These results are admittedly somewhat counter intuitive. They arise from the fact that 10th grade achievement is the product of gains in achievement accumulated over a ten-year period.[7] The noise introduced by this type of aggregation is inevitable if school performance is at all variable over time. (The interested reader may want to peruse appendix tables A-3 and A-4. These tables provide additional information concerning the two simulations discussed above.)[8]

The problem of aggregation over time and grade levels also introduces noise into the comparisons of different schools at the same point in time. The degree to which noise of this type affects the relative ranking of schools depends on whether the variance over time in average achievement growth is large relative to the variance across schools in achievement growth. To illustrate this point, figure 3 considers the consequences of aggregation over time and grade levels for two schools that are identical in terms of average achievement gains over the long term. In the short term, however, average achievement gains are assumed to vary cyclically. For school 1, average gains alternate between ten years of gradual decline and ten years of gradual recovery. For school 2, average gains alternate between ten years of gradual improvement and ten years of gradual decline. These patterns are depicted in panel B of figure 3. The *correct* ranking of schools, based on average achievement growth, is noted in the graph. Panel A depicts the associated levels of average tenth-grade achievement for the two schools. The ranking of schools based on this indicator is also noted. The striking aspect of figure 3 is that the average tenth-grade test score ranks the two schools incorrectly exactly 50 percent of the time. In short, the noise introduced by aggregation over time and grade levels is

Figure 3. Average Tenth-Grade Achievement Given Alternative Cycles of Decline and Recovery in Average Gain



Note: Highest ranked school, according to this indicator, is indicated by the number in the black square.

Note: The correct ranking of schools, as determined by average gain, is indicated by the number in the black square.

particularly troublesome if one is comparing schools that are roughly comparable in terms of long-term average achievement growth. On the other hand, this problem is less serious for schools that differ dramatically in terms of long-term average achievement growth. It is also less serious if cycles of decline and improvement tend to be perfectly correlated. This seems unlikely as a general rule.

The third and final set of simulations illustrates the possible consequences of contamination due to student mobility. These simulations illustrate the extreme sensitivity of average test scores to in-migration of students. To highlight the consequences of student mobility I assume that achievement gains within a school are identical at all grade levels and over time. The first simulation envisions an environment in which there are three types of schools that vary in terms of their average achievement growth.[9] Panel A of table 3 reports the effects on average 10th grade achievement of alternative rates of student mobility among the three schools. Panel B of table 3 reports the fraction of students who change schools, given alternative annual rates of student mobility. Notice that student mobility causes average tenth-grade test scores to collapse toward zero, the average level. For the high- and low-gain schools, for example, an annual mobility rate of 20 percent leads to a reduction in average test scores of over 70 percent. In other words, the average test score is severely biased against high gain schools that happen to serve highly mobile student populations. These numbers suggest that average test scores are apt to be highly misleading indicators of school quality for schools exposed to high rates of student mobility.[10]

If rates and patterns of student mobility vary over time, average test scores are also apt to provide a misleading picture of actual changes in school quality over time. This point is illustrated in figure 4, which simulates the effects on average tenth-grade achievement of an influx of students from a low-quality to a high-quality school system. Events of this kind undoubtedly occur frequently throughout the nation as school systems merge, communities grow, and the occupational structure of jobs evolve in a local labor market. Panel A of figure 4 simulates the effects of a gradual influx of students that takes place over a ten-year period: 1992–2001. Panel B simulates the effects of an instant influx of students in 1992. Despite the fact that average achievement growth remains constant after the influx of students, average achievement levels decline precipitously following the influx of students under either scenario. In the case of the gradual influx of students, the average level of achievement declines by as much as 50 percent. Moreover, average achievement does not return to its 1991 level until the year 2010. In the case of the instant influx of students, the average level of achievement falls instantly by 90 percent and is back to its 1991 level within a decade. In short, idiosyncratic shifts in patterns of student mobility have the potential to grossly contaminate the average test score as an indicator of contemporaneous school performance.

Table 3. Consequences of Student Mobility

A. Average Tenth-Grade Achievement by School, Given Alternative Rates of Student Mobility

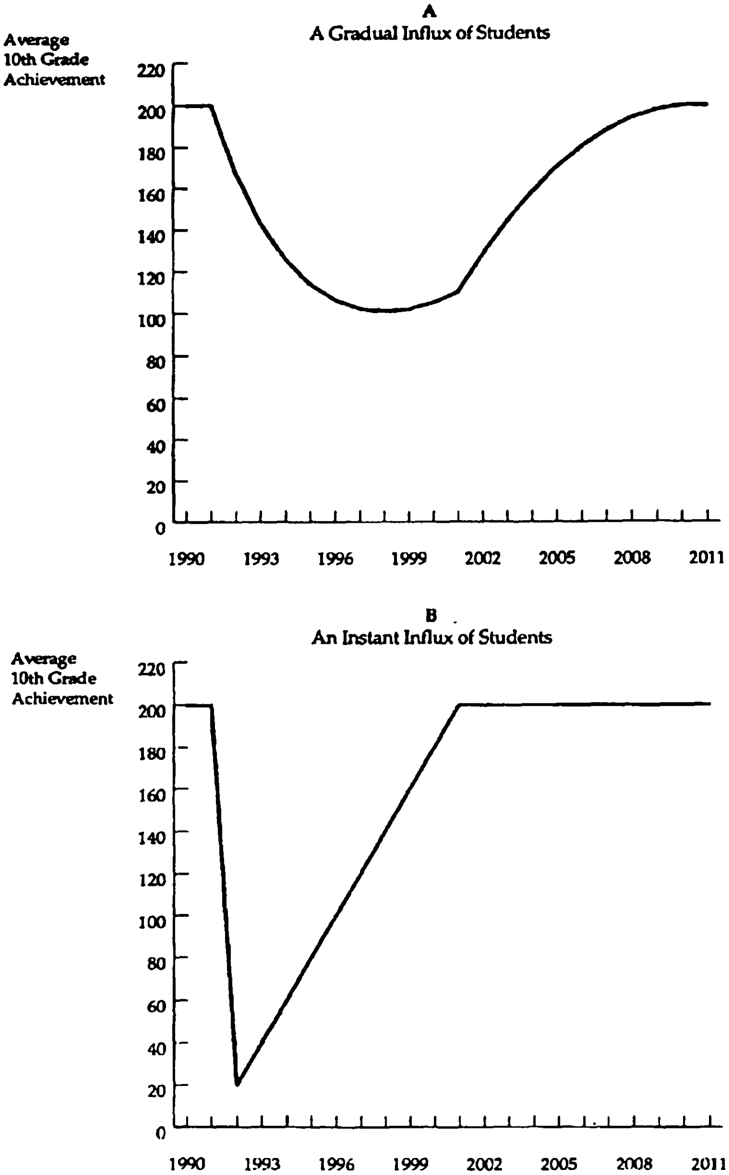
	Annual mobility rate (percent)					
Gain value	**0**	**2**	**5**	**10**	**20**	**40**
High	100	86.9	74.9	56.7	26.8	13.4
Medium	0	4.4	3.4	0.3	-3.8	-2.3
Low	-100	-86.7	-69.3	-56.7	-28.1	-11.3

B. The Fraction of Students Who Change Schools while in Grades 1 through 10, Given Alternative Rates of Student Mobility (percent)

One or more changes	0	17.0	37.0	62.7	89.3	99.7
Two or more changes	0	1.7	8.7	21.7	56.0	94.0

The simulations presented in this section demonstrate that average test scores have the potential to provide a totally misleading portrait of educational productivity, both over time and across schools. Indeed, the simulations possibly understate the degree to which average test scores are flawed as valid measures of school performance since they address the problems of nonlocalization, aggregation, and contamination one at a time, not simultaneously. Fortunately, gain indicators largely avoid the three problems investigated in the above simulations. Moreover, these indicators are generally easy to compute.

Figure 4. Average Tenth-Grade Achievement Given Different Patterns of Student Mobility



A
A Gradual Influx of Students

B
An Instant Influx of Students

An Example Based On National Data

The policy significance of the above discussion is aptly illustrated using data on average mathematics scores from 1973 to 1986 from the National Assessment of Education Progress (NAEP). As indicated in panel A of table 4, NAEP scores for eleventh graders exhibit the by-now familiar pattern of sharp declines from 1973 to 1982 and then partial recovery between 1982 to 1986. The eleventh-grade data, by themselves, are fully consistent with the premise that academic reforms in the early and mid-1980s generated substantial gains in academic achievement. In fact, an analysis of the data based on gain indicators rather than average test scores suggests the opposite conclusion—see panel B of table 4. Gain indicators were constructed in panel B by computing the change in average test scores over time for given birth cohorts.[11] The gain indicators reveal that achievement growth during the 1982 and 1986 period was actually no better than achievement growth during the prior 1978 to 1982 period. In fact, gains from seventh to eleventh grade were actually slightly lower during the 1982 to 1986 period than in previous periods! The rise in eleventh-grade math scores from 1982 to 1986 apparently stems from an earlier increase in achievement growth for that cohort rather than from an increase in achievement growth over grades seven to eleven. In short, these data provide no support for the notion that high school academic reforms generated significant increases in test scores during the mid-1980s. These data also vividly confirm the general superiority of gain indicators, relative to level indicators, as measures of educational productivity.

Value-Added Indicators

As discussed in the previous section, the gain indicator measures the joint contribution of students, families, communities, and schools to growth in student achievement. The problem is that a school may rate highly in terms of a gain indicator primarily or solely because the school serves students capable of rapid achievement growth. Unfortunately, failure to achieve a valid measurement of school performance could provide schools with the incentive to improve "measured" performance simply by trying to control the types of students who attend their schools.

Table 4. NAEP Mathematics Exam Data

A. Average Test Scores				
Grade/Age	1973	1978	1982	1986
3rd/9	219.1	218.6	219.0	221.7
7th/13	266.0	264.1	268.6	269.0
11th/17	304.4	300.4	298.5	302.0

B. Average Test Score Gains			
	1973 to 1978	1978 to 1982	1982 to 1986
3rd to 7th/9 to 13	45.0	50.0	50.0
7th to 11th/13 to 17	34.4	34.4	33.4

SOURCE: Dossey et al (1988).

In order to isolate the distinct contribution of a school to growth in student achievement, a statistical model must be used. The statistical model, if valid, allows one to estimate for each school or classroom the expected (or average) gain in achievement that would be realized by a given student. In this sense, the model estimates school performance *controlling* for differences across schools in student characteristics and perhaps school-level variables such as aggregate student and community characteristics. If these characteristics differ significantly across schools or classrooms, value-added and gain indicators could differ significantly. Dyer, Linn, and Patton (1969), Hanushek (1972), and Murnane (1975) were among the first researchers to estimate value-added indicators of school performance.

What variables should be included as control variables in a value-added model of student achievement and school performance? From the perspective of school accountability, it is important to control for all factors external to schools, in particular, student and community characteristics. Performance with respect to intrinsic school and classroom factors is what matters. In practice, most school districts have ready access to some, but not all, of the student characteristics that are likely to determine student achievement: (1) Is a student eligible for a free or reduced-price school lunch? (2) Is a student eligible for special education services? (3) Does a student's family receive financial assistance from welfare programs? and (4) Is the student classified as being at-risk? It is not well known whether these variables adequately control for differences across schools in average student characteristics. If not, value-added indicators, as implemented, might not fully eliminate the distortions (see below) associated with level and, to a lesser extent, gain indicators.

The exact relationship between a gain and value-added indicator is as follows. For a given cohort at a given grade level, the average gain in student achievement *G* is the sum of two terms: the value-added contribution of a school to growth in student achievement *P* and the average contribution of (external) student and community characteristics to growth in student achievement *F*(*X*), where *X* represents a set of student and community characteristics, and the function F is estimated from an appropriate statistical model of student achievement growth.[12] Similarly, a level indicator is the sum of three terms: *P*, *F*(*X*), and average achievement prior to entering a given grade (see above section on simulation results). From the perspective of measuring school performance, the term *F*(*X*) is a source of error in a gain and level indicator. Prior average achievement is an additional source of error in a level indicator.

The fact that gain and level indicators measure school performance with error has important implications for the use of these indicators for purposes of school choice and accountability. Because of the contamination due to these errors, level indicators, and to a lesser extent gain indicators, are likely to give students the wrong signals about which schools to attend. In practice, this means that prospective students, both academically advantaged and disadvantaged, could be fooled into abandoning an excellent neighborhood school simply because the school served students that were disproportionately academically disadvantaged. At the other extreme, these indicators could contribute to complacency on the part of families whose children attend schools that disproportionately serve academically advantaged students. In fact, these schools could be adding relatively little to the achievement growth of their students. In short, indicators other than the value-added performance indicator convey potentially inaccurate information about school quality and therefore are likely to distort the school choices of students and families. As a result, student achievement is apt to be lower than it would otherwise be.

The consequences of using invalid performance indicators for purposes of public accountability are if anything potentially much worse than in the case of school choice. This stems from the fact that the indicators used for purposes of public accountability have the potential to influence, if not determine, the objectives of teachers and administrators. Indeed, if teachers and administrators are in any way rewarded or penalized on the basis of their performance with respect to a given indicator, they are likely to respond to these incentives by trying to improve their measured performance. In other words, they will have an incentive to "teach to the test." More to the point, they will have an incentive to "teach to the indicator derived from the test."

This phenomenon is the key to understanding why valid performance indicators are potentially capable of generating substantial genuine improvements in school quality. However, it is also the key to understanding how statistically invalid indicators could severely distort the behavior of teachers and administrators. Consider, for example, the consequences of using a level indicator to evaluate school performance. A level indicator is the sum of school performance and two error components that are determined by average student characteristics, average prior achievement, and community characteristics. If this indicator is used to evaluate school performance, it provides teachers and administrators with the incentive to raise measured school performance by teaching only those students who rate highly in terms of average student characteristics, average prior achievement, and community characteristics. In general, these students will be high socioeconomic status, academically advantaged students. This is the phenomenon referred to earlier as "creaming".

The potential for creaming is apt to be particularly strong in environments where schools have the authority to admit or reject prospective students and to expel already enrolled students. However, the problem could also exist in more subtle but no less harmful forms. For example, schools could: (1) create an environment that is relatively inhospitable to academically disadvantaged students, (2) provide course offerings that predominantly address the needs of academically advantaged students, (3) fail to work aggressively to prevent students from dropping out of high school, (4) err on the side of referring "problem" students to alternative schools, (5) err on the side of classifying students as special education students (if these students are exempted from statewide testing), and (6) make it difficult for low-scoring students to participate in statewide examinations. These activities are all designed to improve average test scores in a school, not by improving school quality but by selecting high-scoring students.

As an alternative to trying to select high-scoring students, high-quality teachers and administrators could gravitate to neighborhood schools that predominantly serve high-scoring students. Hence, using the average test score as a high-stakes performance indicator could trigger an exodus of highly skilled educators from schools that disproportionately serve academically disadvantaged students.

One final problem with the average test score is that teachers, administrators, and the public are apt to correctly perceive it as an unfair measure of school performance, thereby undermining the legitimacy of the entire indicator system. Indeed, there is some evidence that this has occurred in one of the states studied by Dominitz and Meyer (1991).

The criticisms discussed above apply equally, although with less force, to the gain indicator, since it is subject to a single source of error, $F(X)$.

Multiple Dimensions Of Performance

Thus far, I have ignored the fact that schools typically have multiple objectives, both academic and nonacademic. Several issues that arise in the context of multiple objectives need to be addressed at this point. First, it seems likely that an ideal performance indicator system would include separate indicators designed to match each and all of the objectives adopted by a school. Such a system would probably include indicators designed to measure school performance in conventional academic subjects, possibly mathematics, science, literature, history, reading, and writing; but it could also include indicators of school performance in other areas, for example, citizenship, employment readiness, and fine arts. The problem is that it could prove technically difficult, burdensome, and expensive to measure outcomes in all of these areas. If indicators are available for only a subset of objectives, however, it is possible, even likely, that those objectives would effectively dominate all other objectives. This could distort the behavior of teachers and administrators by giving them the incentives to devote most of their instructional time to the subjects covered by performance indicators.

One solution to this dilemma is to measure school performance in the areas that are considered to be central to the missions of schools. Indeed, there could be advantages to adopting a more limited set of educational objectives than currently exists. The adoption of performance indicators could conceivably force parents and educators to decide what educational objectives are really important.

It seems inevitable, though, that some important educational objectives could be too difficult to measure. If so, one alternative is to measure the inputs (instructional time and resources) devoted to these activities. This could counteract the incentives to limit instruction in these activities in order to devote more time to activities that are evaluated. On the other hand, the absence of performance indicators in particular areas eliminates the opportunity to hold schools accountable for their performance in these areas.

Second, it seems likely that some educational objectives could be more important than others. How can priorities of this nature be incorporated into an indicator system? One possibility is to construct an overall performance indicator that reflects the preferences of an individual, community, or state. A linear, weighted average of individual performance indicators is one particularly simple example of a preference function. Such a system has recently been adopted in California (Dominitz and Meyer 1991). One potential weakness of the linear preference function is that it allows high performance in one dimension to substitute fully for low performance in another. In fact, most students and parents are likely to prefer schools that are very good in many dimensions, as opposed to schools that are excellent in some areas, poor in others. If so, states and communities could adopt preference functions that limit the degree of substitutability between competing objectives. Examples of such functions include the Cobb-Douglas and constant elasticity of substitution (CES) functions (Henderson and Quandt 1971). This is clearly an area for further research.

Conclusions and Recommendations

The average test score, one of the most commonly used indicators in American education, is highly suspect as an indicator of school performance.[13] This indicator suffers from four major deficiencies: it fails to localize school performance to the classroom or grade level; it aggregates information on school performance across time and grade levels; it is contaminated by student mobility; and it fails to measure the distinct contribution of schools to growth in student achievement. As a result, the average test score is a weak, if not counterproductive, instrument of public accountability. The gain indicator, on the other hand, avoids three of the four problems that plague the average test score. As such, it is a very useful descriptive indicator. The value-added indicator has the major advantage that it avoids all four of the problems that affect level indicators. In particular, it eliminates the incentive for schools to cream.

The value-added approach to measuring school performance relies on a statistical model to identify the distinct contributions made by schools to growth in student achievement. The quality of a value-added indicator is determined by four factors: the frequency with which students are tested, the quality and appropriateness of the tests that underlie the indicators, the adequacy of the control variables included in the appropriate statistical models, and the technical quality of the statistical models used to construct the indicators.

In terms of the first issue, I believe that states need to seriously consider testing students at every grade level, as is currently done in South Carolina (Dominitz and Meyer 1991), or at least at every other grade level, beginning with kindergarten. Annual testing maximizes accountability by localizing school performance to the most natural unit of accountability, the grade level or classroom. It also limits the contamination caused by student mobility and yields up-to-date information on school performance. Less frequent testing, for example, testing at grades kindergarten, four, eight, and twelve, might be acceptable for national purposes, since student mobility is not really at issue at the national level.[14] For purposes of evaluating local school performance, however, the problems created by student mobility argue strongly for frequent testing. To limit the costs and burden imposed by frequent student tests, however, it might be sensible to vary the frequency of testing across schools. Annual testing could be implemented only in schools or school districts where student mobility is high. In addition, annual testing could be implemented in areas with limited enrollments in order to improve the reliability of estimates in these areas, and in schools with low measured performance in order to monitor these schools with greater vigilance.

With respect to the second and third issues, it is important that states make it a major priority to collect extensive and reliable information on student and family characteristics and to develop state tests that are technically sound and fully attuned to their educational goals. Finally, further research is needed to assess the sensitivity of estimates of school performance indicators to alternative statistical models.

NOTES

1. For diagnostic purposes student test scores are often reported separately by subskill areas.

2. This point also applies to classrooms that serve students in more than one grade and ungraded classrooms.

3. This would occur, for example, if the variability over time of school performance is higher in elementary school than in middle or high school.

4. This assumption guarantees that differences across schools in average gain reflect differences in school performance rather than differences in student characteristics.

5. For example, the cohort born in 1980 entered first grade in 1986 and is expected to complete twelfth grade in 1998. Note that the concept of the birth cohort needs to be modified slightly to accommodate school districts that require first graders to be six years old *prior* to beginning school.

6. To further facilitate comparisons across schools at the same grade level, gain values could be standardized so that the standard deviation is equal to ten for every grade in every year. The disadvantage of this approach is that gain indicators constructed in this fashion are not comparable across grades or over time.

7. In the simulations discussed in the text, the average tenth-grade test score is, in fact, exactly equal to a ten-year moving average of average achievement gains. This stems from the simple assumption that achievement gains are identical at different grade levels in the same year.

8. The appendix tables report achievement gains by grade level and cohort. As indicated in the text, achievement gains change from year to year but are always identical across different grade levels in the same year. This shows up in appendix tables A-3 and A-4 as gain values that are equal on diagonal lines that run from the bottom left to the top right of the tables.

9. Average growth is assumed to be equal to 10, 0, and -10, respectively, in the three types of schools. See appendix A for additional details.

10. This conclusion is based on the assumption that at least some student mobility occurs across schools of different quality, a reasonable supposition, we think, in the absence of contrary data.

11. NAEP was originally designed to permit this type of analysis. In mathematics, the tests have generally been given every four years at grade levels spaced four years apart. For this illustrative analysis, I assume that average test scores in 1973 are comparable to the unknown 1974 scores.

12. For concreteness, consider the following statistical mode of achievement growth for students in a given grade:

$$Y(i,j) = a(j) + \sum_{k} b(k)\, X(i,j,k) + e(i,j)$$

where i and j index individuals and schools respectively, Y represents growth in student achievement, X represents a set of student and community characteristics (indexed by k), $a(j)$ represents a school-specific intercept, b represents a set of coefficients (indexed by k), and e represents a random error term. The gain indicator for school j is given by $G = \Sigma\, Y(i,j)/n(j)$, where $n(j)$ = the number of students in school j. The value-added performance indicator for school j is given by $P = a(j)$. The average contribution of external characteristics in school j is given by

$$F(X) = \sum_{i} \sum_{k} b(k)\, X(i,j,k)/n(j).$$

13. Other level indicators, such as the median test score, are similarly suspect.

14. A kindergarten test is needed so that the growth in student achievement in grades one through four can be monitored. In our view, the National Assessment of Educational Progress and recent proposals for national testing in grades four, eight, and twelve are seriously flawed by their failure to include a test at the kindergarten or first-grade level. I suspect that one reason for this omission is that both enterprises are insufficiently aware of the flaws of level indicators and insufficiently aware of the advantages of gain and value-added indicators.

Appendix
Descriptions Of Reported Simulations

This appendix presents results for the simulations presented in the text. Each simulation is defined in terms of the gain in achievement accrued by a student at a particular school in a given grade at a given point in time. The birth cohort subscript is implied by the grade and time subscripts, as discussed in the text. It is given by $c = t - g - 6$. For simplicity, I assume that students begin first grade at age six and advance to subsequent grades one year at a time. Gains in achievement are reported by grade and cohort and tenth-grade achievement for some of the simulations. Gains in achievement for a given year are reported on diagonal lines that run from the bottom left to the top right of the tables.

Appendix Table A-1. Data for Figure 1A

Year cohort completes grade 10	Average gain by grade										Average achievement in grade 10
	1	2	3	4	5	6	7	8	9	10	
1987	0	0	0	0	0	0	0	0	0	0	0
1988	0	0	0	0	0	0	0	0	0	0	0
1989	0	0	0	0	0	0	0	0	0	0	0
1990	0	0	0	0	0	0	0	0	0	0	0
1991	0	0	0	0	0	0	0	0	0	0	0
1992	0	0	0	0	0	0	0	0	0	1	1
1993	0	0	0	0	0	0	0	0	1	2	3
1994	0	0	0	0	0	0	0	1	2	3	6
1995	0	0	0	0	0	0	1	2	3	4	10
1996	0	0	0	0	0	1	2	3	4	5	15
1997	0	0	0	0	1	2	3	4	5	6	21
1998	0	0	0	1	2	3	4	5	6	7	28
1999	0	0	1	2	3	4	5	6	7	8	36
2000	0	1	2	3	4	5	6	7	8	9	45
2001	1	2	3	4	5	6	7	8	9	10	55

Appendix Table A-2. Data for Figure 1B

Year cohort completes grade 10	Average gain by grade										Average achievement in grade 10
	1	2	3	4	5	6	7	8	9	10	
1987	0	0	0	0	0	0	0	0	0	0	0
1988	0	0	0	0	0	0	0	0	0	0	0
1989	0	0	0	0	0	0	0	0	0	0	0
1990	0	0	0	0	0	0	0	0	0	0	0
1991	0	0	0	0	0	0	0	0	0	0	0
1992	0	0	0	0	0	0	0	0	0	2	2
1993	0	0	0	0	0	0	0	0	2	4	6
1994	0	0	0	0	0	0	0	2	4	6	12
1995	0	0	0	0	0	0	2	4	6	8	20
1996	0	0	0	0	0	0	4	6	8	10	28
1997	0	0	0	0	0	0	6	8	10	12	36
1998	0	0	0	0	0	0	8	10	12	14	44
1999	0	0	0	0	0	0	10	12	14	16	52
2000	0	0	0	0	0	0	12	14	16	18	60
2001	0	0	0	0	0	0	14	16	18	20	68

Appendix Table A-3. Data for Figure 2A

Year cohort completes grade 10	Average gain by grade										Average achievement in grade 10
	1	2	3	4	5	6	7	8	9	10	
1981	19	18	17	16	15	14	13	12	11	10	145
1982	18	17	16	15	14	13	12	11	10	9	135
1983	17	16	15	14	13	12	11	10	9	8	125
1984	16	15	14	13	12	11	10	9	8	7	115
1985	15	14	13	12	11	10	9	8	7	6	105
1986	14	13	12	11	10	9	8	7	6	5	95
1987	13	12	11	10	9	8	7	6	5	4	85
1988	12	11	10	9	8	7	6	5	4	3	75
1989	11	10	9	8	7	6	5	4	3	2	65
1990	10	9	8	7	6	5	4	3	2	1	55
1991	9	8	7	6	5	4	3	2	1	0	45
1992	8	7	6	5	4	3	2	1	0	1	37
1993	7	6	5	4	3	2	1	0	1	2	31
1994	6	5	4	3	2	1	0	1	2	3	27
1995	5	4	3	2	1	0	1	2	3	4	25
1996	4	3	2	1	0	1	2	3	4	5	25
1997	3	2	1	0	1	2	3	4	5	6	27
1998	2	1	0	1	2	3	4	5	6	7	31
1999	1	0	1	2	3	4	5	6	7	8	37
2000	0	1	2	3	4	5	6	7	8	9	45
2001	1	2	3	4	5	6	7	8	9	10	55

Appendix Table A-4. Data for Figure 2B

Year cohort completes grade 10	Average gain by grade										Average achievement in grade 10
	1	2	3	4	5	6	7	8	9	10	
1981	19	18	17	16	15	14	13	12	11	10	145
1982	18	17	16	15	14	13	12	11	10	9	135
1983	17	16	15	14	13	12	11	10	9	8	125
1984	16	15	14	13	12	11	10	9	8	7	115
1985	15	14	13	12	11	10	9	8	7	6	105
1986	14	13	12	11	10	9	8	7	6	5	95
1987	13	12	11	10	9	8	7	6	5	6	87
1988	12	11	10	9	8	7	6	5	6	7	81
1989	11	10	9	8	7	6	5	6	7	8	77
1990	10	9	8	7	6	5	6	7	8	9	75
1991	9	8	7	6	5	6	7	8	9	10	75
1992	8	7	6	5	6	7	8	9	10	10	76
1993	7	6	5	6	7	8	9	10	10	10	78
1994	6	5	6	7	8	9	10	10	10	10	81
1995	5	6	7	8	9	10	10	10	10	10	85
1996	6	7	8	9	10	10	10	10	10	10	90
1997	7	8	9	10	10	10	10	10	10	10	94
1998	8	9	10	10	10	10	10	10	10	10	97
1999	9	10	10	10	10	10	10	10	10	10	99
2000	10	10	10	10	10	10	10	10	10	9	100
2001	10	10	10	10	10	10	10	10	10	10	100

Appendix Table A-5. Data for Figure 3, School 1

Year cohort completes grade 10	Average gain by grade										Average achievement in grade 10
	1	2	3	4	5	6	7	8	9	10	
1981	2	4	6	8	10	12	14	16	18	20	110
1982	4	6	8	10	12	14	16	18	20	18	126
1983	6	8	10	12	14	16	18	20	18	16	138
1984	8	10	12	14	16	18	20	18	16	14	146
1985	10	12	14	16	18	20	18	16	14	12	150
1986	12	14	16	18	20	18	16	14	12	10	150
1987	14	16	18	20	18	16	14	12	10	8	146
1988	16	18	20	18	16	14	12	10	8	6	138
1989	18	20	18	16	14	12	10	8	6	4	126
1990	20	18	16	14	12	10	8	6	4	2	110
1991	18	16	14	12	10	8	6	4	2	0	90
1992	16	14	12	10	8	6	4	2	0	2	74
1993	14	12	10	8	6	4	2	0	2	4	62
1994	12	10	8	6	4	2	0	2	4	6	54
1995	10	8	6	4	2	0	2	4	6	8	50
1996	8	6	4	2	0	2	4	6	8	10	50
1997	6	4	2	0	2	4	6	8	10	12	54
1998	4	2	0	2	4	6	8	10	12	14	62
1999	2	0	2	4	6	8	10	12	14	16	74
2000	0	2	4	6	8	10	12	14	16	18	90
2001	2	4	6	8	10	10	14	16	18	20	110

Appendix Table A-6. Data for Figure 3, School 2

Year cohort completes grade 10	Average gain by grade										Average achievement in grade 10
	1	2	3	4	5	6	7	8	9	10	
1981	18	16	14	12	10	8	6	4	2	0	90
1982	16	14	12	10	8	6	4	2	0	2	74
1983	14	12	10	8	6	4	2	0	2	4	62
1984	12	10	8	6	4	2	0	2	4	6	54
1985	10	8	6	4	2	0	2	4	6	8	50
1986	8	6	4	2	0	2	4	6	8	10	50
1987	6	4	2	0	2	4	6	8	10	12	54
1988	4	2	0	2	4	6	8	10	12	14	62
1989	2	0	2	4	6	8	10	12	14	16	74
1990	0	2	4	6	8	10	12	14	16	18	90
1991	2	4	6	8	10	12	14	16	18	20	110
1992	4	6	8	10	12	14	16	18	20	18	126
1993	6	8	10	12	14	16	18	20	18	16	138
1994	8	10	12	14	16	18	20	18	16	14	146
1995	10	12	14	16	18	20	18	16	14	12	150
1996	12	14	16	18	20	18	16	14	12	10	150
1997	14	16	18	20	18	16	14	12	10	8	146
1998	16	18	20	18	16	14	12	10	8	6	138
1999	18	20	18	16	14	12	10	8	6	4	126
2000	20	18	16	14	12	10	8	6	4	2	110
2001	18	16	14	12	10	8	6	4	2	0	90

References

Clune, William H. 1991. "Systemic Educational Policy." Wisconsin Center For Educational Policy, University of Wisconsin-Madison, July.

Dominitz, Jeff, and Robert H. Meyer. 1991. "Educational Performance Indicators and School Report Cards: Lessons From Three States." University of Wisconsin-Madison. Mimeo.

Dossey, John A., et al. 1988. *The Mathematics Report Card: Are We Measuring Up*? Princeton, N.J.: Educational Testing Services.

Dyer, Henry S., Robert L. Linn, and Michael J. Patton. 1969. "A Comparison of Four Methods of Obtaining Discrepancy Measures Based on Observed and Predicted School System Means on Achievement Tests," *American Educational Research Journal* 6, 4 (November): 591–605.

Goldman, Jay P. 1990. "Grading Schools Through Report Cards: Realtors, News Media Collect them For Comparative Purposes," *School Administrator* 47, 6: 26–30.

Hanushek, Eric A. 1972. *Education and Race*. Lexington, MA: D.C. Heath.

Hanushek, Eric A., and Lori Taylor. 1990. "Alternative Assessments of the Performance of Schools," *Journal of Human Resources* 25, 2 (Spring): 179–201.

Henderson, James M., and Richard E. Quandt. 1971. *Microeconomic Theory: A Mathematical Approach*. New York: McGraw-Hill.

Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children*. Cambridge, MA: Ballinger.

Powell, Brian, and Lala Carr Steelman. 1984. "Variations in State SAT Performance: Meaningful or Misleading," *Harvard Educational Review*, 54, no. 4: 389–412.

Smith, Marshall S., and Jennifer O'Day. 1990. "Systemic School Reform." In *The Politics of Curriculum and Testing*. Edited by Susan Fuhrman and Betty Malen. 1990 Yearbook of the Politics and Education Association. London: Taylor and Francis. Pp. 233–67.

U.S. Department of Education. 1971. *America 2000: An Education Strategy*.

Wainer, Howard. 1986. "The SAT as a Social Indicator: A Pretty Bad Idea." In *Drawing Inferences from Self-Selected Samples*. Edited by Howard Wainer. New York: Springer-Verlag. Pp. 7–22.