Institute for Research on Poverty
Discussion Paper no. 1052-94

Educational Performance Indicators: A Critique

Robert H. Meyer
Harris Graduate School of Public Policy Studies
The University of Chicago

Revised
February 1995

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 Dominitz, and Andy Porter and from comments by Ken Bickers, Tony Bryk, John Easton, Mike Wiseman, and participants in the La Follette Institute of Public Affairs workshop.

Abstract

The typical indicators used to assess a school's performance—average and median achievement test scores—are highly flawed. Simulation results indicate that these indicators provide a severely misleading portrait of changes in school performance over time and differences in performance across schools, particularly if students change schools a lot or school performance varies significantly over time. Moreover, these indicators provide schools with the incentive to cater to students who score high on achievement tests, and they tend to be biased against schools that serve a large number of academically disadvantaged students. Better than average and median test scores are gain indicators, which measure the growth in achievement from one grade to the next for a given group of students, and value-added indicators, which rely on a statistical model to identify the distinct contributions of schools to growth in student achievement at a given grade level. In order to implement valid school performance indicators, schools should test students every two years, if not annually, beginning with kindergarten; collect better data on student and family characteristics; and develop tests that are sound and attuned to their educational goals.

Educational Performance Indicators: A Critique

I. INTRODUCTION

Educational outcome indicators increasingly are being used to assess the efficacy of American education. Local newspapers regularly report how students in local schools perform on nationally standardized tests; a growing number of states publish state and local school report cards that provide an assortment of student outcome, enrollment, and financial indicators; and the federal government, with the support of the nation's governors, has dramatically expanded the nation's testing program.

The growth of educational outcome indicators has been motivated in large part by a growing demand to hold schools accountable for their performance, defined in terms of outcomes, such as standardized test scores, rather than inputs, such as teacher qualifications, class size, and the number of books in school libraries. States that have dramatically increased expenditures on education or have launched major school improvement efforts have been particularly likely to adopt extensive indicator systems. The increased demand for public accountability in elementary and secondary education parallels similar demands for increased accountability in other public sector activities, for example, the Job Training Partnership Act (JTPA) and the JOBS program, enacted as part of the Family Support Act.

Despite the groundswell of interest in data on school performance, many educators and scholars fear that a poorly implemented performance indicator system could ultimately be worse than no indicators at all. These fears are not groundless. Performance indicators based on achievement tests could be flawed in three major ways. First, the achievement test underlying a performance indicator could fail to reflect a school's true educational objectives or otherwise lack validity. As is well known, many educators believe that the vast majority of tests, particularly standardized, multiple choice tests, currently exhibit this flaw. In particular, there is concern that these tests focus almost exclusively on low level academic content and have little to do with schools' educational goals.

(Smith and O'Day, 1990; Clune, 1991). As a result, many school districts, states, and professional test developers are currently experimenting with new types of assessments, including tests with open-ended questions, performance-based assessments, and graded portfolios.[1] Second, a performance indicator could be susceptible to "corruption." For example, a test that is administered year after year could be corrupted if instructors teach narrowly to the test, as opposed to the content domain that underlies the test.[2] Finally, a performance indicator constructed from a simplistic or otherwise inappropriate statistical model could fail to measure the true contribution of a school to growth in measured student achievement. If any one of these flaws exists in a high-stakes system of educational performance indicators, the system could severely distort the behavior of educators and students.

The purpose of this paper is to investigate the third factor, the statistical adequacy of the most commonly used educational performance indicators. Several major conclusions emerge from the analysis. First, the typical indicators used to assess school performance—average and median test scores—are highly flawed as measures of school performance, even if they are derived from highly valid assessments. As a result, they are of limited value, if not useless, for evaluating relative school performance or school performance over time and thus should not be used to hold schools accountable for their performance. Indeed, simulation results indicate that changes over time in average test scores could very well be *negatively* correlated with actual changes in school performance.

Second, 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 that 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

---

[1]See, for example, Wiggins (1989), Darling-Hammond (1991), Shepard (1991), and Koretz et al. (1994).

[2]See, for example, Haladyna, Nolen, and Haas (1991), Nolen, Haladyna, and Haas (1992), Smith and Rottenberg (1991), and Shepard (1991).

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 of or transfer to another school. Second, schools could aggressively retain students at given grade levels (Shepard, 1991). Finally, high-quality teachers and administrators could gravitate to neighborhood schools that predominantly serve high-scoring students.

Third, typical school performance indicators tend to be biased against schools that disproportionately serve academically disadvantaged students. One source of bias is the well-known fact that school productivity is only one of the many determinants of student achievement. Most of the variation in average or median test scores can usually be accounted for by differences across schools in the types of students enrolled. A second source of bias is that school-performance targets, specified in terms of average or median test scores, are likely to be much lower than properly specified (value-added) measures of school performance for schools that disproportionately serve academically disadvantaged students. This has the obvious effect of understating the effort and perhaps resources required to raise schools of this type up to a target level of student achievement.

Finally, given the problem of student mobility (as well as several other problems discussed in the paper), it is not possible to construct statistically valid school performance indicators if tests, assessments, or other student outcomes are measured so infrequently that a significant proportion of students change schools in between periods of testing. In particular, the simulation results reported in the paper suggest that it is simply not enough to test students every four years (for example, in grades 4, 8, and 12), as is currently done in the National Assessment of Education Progress and some state testing programs. This may be okay at the national level, since student mobility in and out of the

country is limited. For purposes of constructing school, district, and perhaps state indicators, however, it appears to be necessary to test with much greater frequency.[3]

Given the substantial problems that exist with common educational indicators, what should be done to improve the situation? If one is interested in having indicators that are appropriate for accountability purposes, I believe that the only solution is to design an indicator system that avoids the three major flaws listed at the outset of this paper. There are reasons to be optimistic that it is possible to do this. A number of states have recently made great strides in designing new tests and assessments that are specifically aligned with state educational goals and appear to avoid many of the problems of multiple-choice tests. From the statistical perspective, the performance indicators implemented in South Carolina, Tennessee, and Dallas, Texas, appear to be particularly promising. They measure school performance using value-added measures of school performance. The other option is to give up on the idea of using outcome indicators to hold schools accountable for their performance. As stated earlier, this option could very well dominate the option of retaining an indicator system that was not originally designed to drive a high-stakes accountability system. If policymakers are serious about holding schools and districts accountable for their contribution to growth in student achievement, it is crucial to measure that contribution validly using student outcome measures that accurately reflect schools' educational goals.

The paper is organized in seven major sections, including the introduction, Section I. Section II presents a simple value-added model of student achievement growth that provides a standard against which alternative indicators of school performance can be evaluated. Section III draws on a series of simulations to demonstrate that the average test score is highly flawed as an indicator of school performance. Section IV draws on data from the National Assessment of Educational Progress to

---

[3]The need for frequent testing lends further support to the idea that tests and assessments need to be closely aligned to educational goals and, if possible, to add to, rather than distract from, the educational process (Clune, 1991; Smith and O'Day, 1990).

illustrate the problems that arise from interpreting the average test score as an indicator of school performance. Section V evaluates the consequences of using flawed indicators as the foundation of a school accountability system. Section VI considers the data required to construct valid and reliable value-added indicators. Section VII addresses the issue of whether value-added indicators lead to higher or lower performance expectations for schools that disproportionately serve disadvantaged students. Finally, Section VIII concludes the paper. Two appendices provide technical information on the simulations reported in section III.

II. CRITERIA FOR EVALUATING THE STATISTICAL VALIDITY OF PERFORMANCE INDICATORS

Standardized student testing is conducted for a variety of reasons: to provide information on individual students and to provide 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, to guide teachers in providing instruction that matches the needs of individual students, to guide students in making curriculum and career choices, to determine whether students are eligible for graduation (in states that have minimum competency examinations), 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, the average level of achievement, or the median level of achievement. I refer generally to statistics of this kind as *level* indicators. Level indicators are widely reported by schools and states. 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 (NAEP). 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 that 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 will limit my analysis to level indicators that are not subject to these problems.

Level indicators, if correctly constructed and based on appropriate tests or assessments, 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 this paper is whether such indicators are adequate for purposes of holding schools accountable for their performance. To do so they must validly measure school or classroom performance, in particular, the contribution of schools to growth in student achievement for students in particular grades or sequence of grades. As will be demonstrated in this paper, level indicators fail to do this, often by a huge margin.

In order to evaluate the validity of alternative school performance indicators it is necessary to specify the benchmark indicators that will be used as the standard of comparison for all other indicators. Consistent with the vast literature on the determinants of achievement growth, I define true school performance using a statistical, value-added model (see, for example, Dyer, Linn, and Patton, 1969; Hanushek, 1972, Murnane, 1975; Raudenbush and Bryk, 1986; Willms and Raudenbush, 1989; Hanushek and Taylor, 1990; and Meyer, 1992). Since the primary objective of this paper is to assess the statistical validity of common outcome indicators, I assume that achievement growth from one grade to the next can be adequately characterized by a two-level, linear growth model (Willett, 1988): where i indexes individual students, s indexes schools, g indexes grade levels, and t indexes school years; Y represents student achievement for a given individual in grade g in year t; X(i) represents a

$$Y(i,g,t) = Y(i,g-1,t-1) + \beta(g)X(i) + \sum_{s} [\alpha(s,g,t) + \gamma(g)\ C(s,g,t)]\ S(i,s,g,t) + e(i,g,t)$$

(1)

set (vector) of individual and family characteristics, assumed (for simplicity) to be invariant over time; S(i,s,g,t) is an indicator variable equal to one if student i is enrolled in school s in grade g in year t, zero otherwise; C(s,g,t) represents a set (vector) of community characteristics and school-aggregate student characteristics (for example, average socioeconomic status); $\beta(g)$ and $\gamma(g)$ are parameters (vectors) that capture the effects of X and C on growth in student achievement; $\alpha(s,g,t)$ is a school effect; and e(i,g,t) is a random component of student achievement growth assumed to be uncorrelated with all regressors included in the model.[4]

The model has a straightforward interpretation: student achievement in a given grade in a given year is equal to student achievement in the prior year and grade, plus a term $\beta(g)X(i)$ that reflects the contribution of individual and family characteristics to growth in student achievement, a term (in brackets in the second line) that reflects the growth in student achievement as a result of attending a given school, and a random term. The effect of attending a given school can further be broken down into two parts: a component $\alpha(s,g,t)$ that is the result of differences in school policies,

[4]This model is a reasonable one if the pre- and post-test scores are scaled so that achievement is measured in the same units. If this is not the case, the model could be extended to allow the pre-test variable to have its own coefficient, possibly different from the value of one that is imposed in the above model. This model has often been used in previous studies. However, Meyer (1992) demonstrates that in a model of this type it is necessary to correct for measurement error in the pre-test variable. Note that the model is defined only for students who attend a given school for the entire school year and have achievement test data both prior to and at the end of the school year. Students who fail to meet these conditions must be excluded from the analysis. In principle, this problem could be avoided by testing students more than once a year, although this would be an expensive and burdensome proposition.

teacher quality, etc., and a component that captures the contribution of community and school-aggregate student characteristics to school effectiveness.[5]

The former factor $\alpha(s,g,t)$ measures the contribution of a school to growth in student achievement after controlling for all factors that are *external* to the school. I refer to this indicator as a measure of *intrinsic* school performance.[6] Willms and Raudenbush (1989) refer to this indicator as a Type B indicator. This indicator can be interpreted as a measure of the collective performance of school staff (at a given grade level) and thus is the indicator that is appropriate for purposes of school accountability. A second value-added indicator, a measure of *total* school performance, is given by $\alpha(s,g,t) + \gamma(g)\, C(s,g,t)$. Willms and Raudenbush refer to this indicator as a Type A indicator. This indicator reflects the intrinsic performance of a school (α) plus the part of school performance that is determined by factors external to the school (community, and school-aggregate student characteristics).[7] One interpretation of this indicator is that it captures the effect of enrolling one additional student in a school, holding community characteristics and the composition of the student group approximately fixed. If these characteristics are relatively stable from year to year, the total school performance indicator is appropriate for purposes of informing school choice. In this paper I focus primarily on the intrinsic school performance indicator.

The value-added indicators derived from equation (1) define school performance at a specific grade level at a particular point in time. The average test score, on the other hand, reflects the

---

[5]Bryk and Raudenbush (1992) discuss methods for estimating models of this type.

[6]The intrinsic school performance indicator is implicitly defined by the school-level control variables (C) included in the model. At a minimum, school-level measures of student and neighborhood characteristics should be included in the model since these variables are determined externally to the school (at least in the short term). The model could also include variables such as per pupil expenditures and the quality of building facilities in order to control for school inputs that may not be controllable by principals and teachers.

[7]See Gamoran (1992) for a survey of research on the effects of school-aggregate student characteristics on school performance.

contribution of school, family, and community inputs at *all grades* prior to the year students are tested. In order to be able to assess the validity of the average test score as a measure of school performance it is necessary to aggregate the grade-specific indicators to obtain an aggregate indicator that measures performance over a sequence of grades, say, 4th through 8th grade or 1st through 10th grade. An aggregate indicator that does this is derived below. I refer to it as a steady state indicator.

The steady state indicator can be motivated in the following way. Imagine that a given school enters a steady state in year t such that the productivity of the school at all grade levels stays constant. For a given group of students who attend school s from grade g_1 to g_2, achievement in grade g_2 is then given by[8]

$$Y(i,g_2,t) = Y(i,g_1-1,t) + [\beta(g_1) + \ldots + \beta(g_2)]X(i) + \{\gamma(g_1)C(s,g_1,t) + \ldots + \gamma(g_2)C(s,g_2,t)\} + SS(s,g_1,g_2,t) + \{e(i,g_1,t) + \ldots + e(i,g_2,t)\},$$

(2)

where the steady state indicator $SS(s,g_1,g_2,t)$ is given by[9]

$$SS(s,g_1,g_2,t) = \alpha(s,g_1,t) + \ldots + \alpha(s,g_2,t).$$

(3)

This equation is similar in form to equation (1): student achievement in grade g_2 is equal to student achievement prior to grade g_1, the cumulative contribution of individual, family, and community characteristics to growth in student achievement over grades g_1 to g_2, the steady state contribution of

---

[8]This equation is derived by successively substituting the equation for $Y(g_2\text{-}1,t)$ into the equation for $Y(g_2,t)$, etc.

[9]A closely related indicator is steady state performance per grade

$$\overline{SS}(s,g_1,g_2,t) = SS(s,g_1,g_2,t)/(g_2 - g_1).$$

This indicator is appropriate for comparing the performance of schools that may not have the same number of grades.

the school over grades g_1 to g_2, and the cumulative contribution of random components of growth. (A steady state indicator of total school performance would also include the terms in the second line of (2).) The steady state indicator defined above is intuitively quite plausible in the context of the model of achievement growth adopted in this paper. It is simply the sum of school performance indicators in year t over the specified range of grades.[10] As will be demonstrated below, this indicator is potentially quite different from a school-level average test score.

Given the above model of student achievement growth, it is possible to define and evaluate all of the outcome indicators commonly used to evaluate schools. In this paper I focus on the most common of all outcome indicators, the average test score, and a simple alternative to that indicator, the gain indicator. In order to define these indicators in terms of the parameters of the value-added model it is necessary to introduce some additional notation. For any variable, a bar over the variable name denotes the school-level mean for the group of students enrolled in school s in grade g in year t. For example, the school-level average of X(i) in school s in grade g and year t is given by $\overline{X}(s,g,t)$ and the school-level average of S(i, s', g', t') is given by $\overline{S}(s',g',t' \mid s,g,t)$. The latter variable is simply the fraction of students (since S is a zero/one indicator) in school s in grade g in year t who previously were enrolled in school s' in grade g' in year t'.

Given this notation the gain indicator is defined by

$$G(s,g,t) \equiv \overline{Y}(g,t \mid s,g,t) - \overline{Y}(g-1,t-1 \mid s,g,t), \quad \textbf{(4)}$$

[10]The steady state indicator presented in this paper can be adapted to handle a wide range of alternative achievement growth models. The specific formula for the steady state indicator depends on the exact structure of the value-added model.

that is, the growth in average achievement from one grade to the next for a given cohort of students.[11] This indicator, in general, is very different from the change over time in average achievement at a given grade level, as is indicated below. Given the assumed model of achievement growth, the gain indicator is given by[12]

$$G(s,g,t) = \alpha(s,g,t) + \text{Adjustment Factor } (s,g,t) \quad \textbf{(5)}$$

where

$$\text{Adjustment Factor } (s,g,t) = \beta(g)\overline{X}(s,g,t) + \gamma(g)C(s,g,t).$$

In other words, the average gain in student achievement is equal to intrinsic school performance plus an adjustment factor that reflects the aggregate, school-level effects of student, family, and community characteristics on student achievement growth. It is immediately apparent that the gain indicator, taken as a measure of intrinsic school performance, overstates the performance of schools that disproportionately serve students and communities that are academically advantaged. The opposite is true for schools that disproportionately serve students and communities that are academically disadvantaged. The gain indicator is thus biased against schools of the latter type, if it is interpreted as a measure of school performance. (The criticism also applies to the average test score, as discussed below). The effects on schools of using this indicator for accountability purposes are explored later in the paper.

In order to derive an appropriate formula for the average test score, equation (1) must be rewritten so as to eliminate the prior test score variables for all grades beyond the initial one, grade

---

[11]The gain indicator has a meaningful interpretation only if the post- and pre-test scores are scaled so that achievement is measured in the same units.

[12]For simplicity, the average error $\overline{e}(g,t,|s,g,t)$ is excluded from this equation.

0.[13] This yields a reduced form model (see Boardman and Murnane, 1979) that expresses student achievement as the outcome of initial achievement, the contribution of student, community, and intrinsic school effects for all prior grades, and the sum of all prior individual error terms. The average test score for students in school s in grade 10, say, in year t is then given by[14]

$$\begin{aligned}\bar{Y}(10,t\,|\,s,10,t) &= \bar{Y}(0,t-10\,|\,s,10,t) + [\beta(1) + \ldots + \beta(10)]\ \bar{X}(s,10,t) \\ &+ \{\sum_{s'} \gamma(1)C(s',1,t-9)\,\bar{S}(s',1,t-9\,|\,s,10,t) + \\ &\ldots + \sum_{s'} \gamma(10)\,C(s',10,t)\,\bar{S}(s',10,t\,|\,s,10,t)\} \\ &+ \{\sum_{s'} \alpha(s',1,t-9)\,\bar{S}(S',1,t-9\,|\,s,10,t) + \ldots + \sum_{s'} \alpha(s',10,t)\,\bar{S}(s',10,t\,|\,s,10,t)\}\end{aligned} \qquad \textbf{(6)}$$

In other words, average 10th grade achievement for a given group of students in school s in year t is the sum of three terms (on the first three lines) that reflect differences across schools in student, family, and community characteristics[15] and a term (on the last line) that reflects the intrinsic school performances of *all schools* attended by the given group of students in grades 1 through 10. The latter term, although similar to the steady state indicator, differs from it in two important respects. First, it reflects the intrinsic performance of all schools attended by students in grades 1 through 10 (not just

[13]As in the case of the steady state indicator, this equation is derived by successively substituting the equation for $Y(g_2\text{-}1, t\text{-}1)$ into the equation for $Y(g_2, t)$, etc.

[14]Again, for simplicity, the average error term is excluded from this equation.

[15]The three terms represent the average initial achievement of the students prior to entering 1st grade, the effect of student-level individual and family characteristics, and the effect of school-level characteristics at *all schools* attended by the group of students in grades one through ten.

the school in which the students are enrolled in grade 10).[16] Second, it reflects the intrinsic performance of schools over a ten-year period (not just the current year).

In the next section I draw on equation (6) to conduct a detailed comparison of the difference between average test scores and intrinsic school performance, measured at each grade level and as a steady state aggregate.

III. A CRITIQUE OF THE AVERAGE TEST SCORE AS A MEASURE OF SCHOOL PERFORMANCE: THEORY AND SIMULATION RESULTS

The average test score is highly flawed as a measure of school performance for four basic reasons. One, as in the case of the gain indicator, the average test score is *contaminated by factors other than school performance,* in particular, the average level of student achievement prior to entering 1st grade—average initial achievement—and the average effects of student, family, and community characteristics on student achievement growth from 1st grade through the grade in which students are tested. (These factors are given in the first three lines of equation (6)). In fact, it is quite likely that comparisons across schools of average test scores primarily reflect these differences rather than genuine differences in intrinsic school performance. As such, average test scores are highly biased against schools that disproportionately serve academically disadvantaged students and communities.

Two, the average test score reflects information about school performance that tends to be grossly *out of date*. For example, consider the average test score for a group of 10th grade students. The test scores for these students reflect learning that occurred in kindergarten, roughly 10½ years

---

[16]In the 1st grade, for example, the contribution of each school (α(s', 1, t-9)) to the average test score for school s is weighted by

$$\bar{S}\,(s', 1, t-9 \mid s, 10, t),$$

the fraction of students in school s who attended school s' in 1st grade (nine years earlier).

earlier, through the 10th grade. Indeed, a 10th grade level indicator could be dominated by information that is five or more years old.[17] 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 presumably must reflect information that is current.

Three, average test scores at the school, district, and state levels tend to be highly *contaminated due to student mobility* in and out of different schools. For example, the typical high school student is likely to attend several different schools over the period spanning kindergarten through 12th grade. For these students, 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 may differ dramatically across schools. Contamination is apt to be especially high in communities that undergo rapid population growth or decline and in communities that 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, but at the state, district, and school levels it is apt to be quite serious.

Finally, unlike the grade-specific value-added indicator, 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. This lack of localization is, of course, most severe at the highest grade levels. 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.

---

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

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:

- failure to localize school performance to specific grade levels,
- aggregation of performance information that is grossly out of date,
- contamination due to student mobility,
- contamination due to nonschool factors (students, family, and community characteristics).

The latter flaw is also shared by the gain indicator.

Below, I present a series of simulations that illustrate how these factors have the potential to dramatically distort the average test score as a measure of school performance. The simulations illustrate the consequences of each factor separately. In practice, of course, all four problems are likely to coexist.

The simulations are designed to assess the validity of the average test score with respect to two applications: (1) comparisons of indicators across schools and (2) comparisons of indicators over time for the same school. The latter type of comparison is particularly relevant for the purposes of evaluating the efficacy of school reform efforts.

In the first three sets of simulations reported in this section, I assume that average initial achievement and average student characteristics are identical for all schools at all points in time.[18] Hence, the analysis focuses on problems created by the last line of equation (6). Given alternative assumptions concerning the pattern of intrinsic school performances over time and across grade levels, I compute the average level of achievement at the end of grade 10 using equation (6) and the steady state indicator for grades 1 through 10 using equation (3). In all but one of the simulations I assume,

[18]I also assume that the average error at every school at every point in time is zero. I invoke this assumption so that I can ignore the issue of reliability and focus entirely on the validity—or lack of validity—of indicators.

for simplicity, that school performance is identical at all grade levels in a given year. For these simulations the steady state indicator is equal to ten times the grade-level indicator of school performance. To facilitate comparisons across schools, I standardize the intrinsic school performance values so that they are approximately centered around zero and range from -20 to 20. The technical details of the simulations are presented in Appendices A and B. Appendix A also includes all of the simulation data reported in the graphs discussed in the text.

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 aggregation of out-of-date information, student mobility, and differences across schools in student, family, and community characteristics.

The first simulation, as summarized in Table 1, contrasts three schools that differ in terms of their patterns of school performance in grades 1 through 5 and grades 6 through 10, respectively. To simplify the analysis I assume that these patterns persist over time and that there is no student mobility. School 1 exhibits school-performance values of 0 (the average) at all grade levels. School 2 exhibits exceptionally high school performance values in the lower grades and exceptionally low school-performance values in the higher grades. Finally, school 3 exhibits a pattern of school performance values that is exactly opposite to the pattern exhibited for school 2. As indicated, the three schools differ fundamentally in terms of their school performance 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 10th grade. The exceptionally high and the exceptionally low performance 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 10th 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

TABLE 1

**Average 10th Grade Achievement by School,
Given Alternative Patterns of Intrinsic School Performance**

School Grade	Initial Achievement	Intrinsic School Performance by Grade: Grades 1 to 5	Intrinsic School Performance by Grade: Grades 6 to 10	Achievement at the End of Tenth
1	0	0	0	0
2	0	20	-20	0
3	0	-20	20	0

Source: Data simulated by author.

Figure 1 here

stable, but average, school performance at all grade levels. Panels A and B in Figure 1 depict two different scenarios. In Panel A school performance at each grade level increases gradually after 1991. In Panel B, school performance also increases steadily, but the improvement is limited to grades 7 to 10. As in the previous simulation, the two schools differ substantially in terms of their school performance values at different grade levels. Despite these differences, however, there is no perceptible difference between the two schools in terms of average 10th 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.

The second set of simulations illustrates the problem of school performance information that tends to be grossly out of date. 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 achievement gains. To highlight the problem of aggregation across time and grade levels I assume that school performance within a school is identical at all grade levels and that there is no student mobility. Figure 2 charts average 10th grade achievement and school performance 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 school performance across all grades and initiate a trend of gradual improvement in school performance across all grades. Panel B of Figure 2 depicts a scenario in which academic reforms have absolutely no effect on school performance. The reforms, however, are preceded by an era of gradual deterioration in school performance across all grades, followed by a brief period (1987 to 1991) of gradual improvement across all grades. As indicated in the graph, the average 10th grade test score provides a totally misleading view of the effectiveness of the hypothetical academic reforms

Figure 2 here

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 10th 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 counterintuitive. They arise from the fact that 10th grade achievement is the product of gains in achievement accumulated over a ten-year period.[19] 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.)[20]

The problem of aggregation of information that is grossly out of date 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 school performance over the long term. In the short term, however, school performance is assumed to vary cyclically. For school 1, school performance alternates between ten years of gradual decline and ten years of gradual recovery. For school 2, school performance alternates 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 school performance, is noted in the graph. Panel A depicts the associated levels of average 10th

[19]In the simulations discussed in the text, the average 10th grade test score is, in fact, exactly equal to a ten year moving average of school performance. This stems from the simple assumption that school performance is identical at different grade levels in the same year.

[20]The tables in Appendix A report school performance by grade level and cohort. As indicated in the text, school performance changes from year to year but is always identical across different grade levels in the same year. This shows up in Appendix Tables A-3 and A-4 as school performance values that are equal on diagonal lines that run from the bottom left to the top right of the tables.

Figure 3 here

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 10th grade test score ranks the two schools correctly only 50 percent of the time. In short, the noise introduced by aggregation over time and grade levels is particularly troublesome if one is comparing schools that are roughly comparable in terms of long-term school performance. On the other hand, this problem is less serious for schools that differ dramatically in terms of long-term average school performance. It is also less serious if cycles of decline and improvement tend to be perfectly correlated across schools. This seems unlikely as a general rule.

The third 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 school performance within a school is 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 school performance.[21] Student mobility is assumed to follow a Markov process: at the end of each school year all students either stay in their current school or move, with some probability, to one of the other two schools. The technical details of this simulation are reported in Appendix B. Panel A of Table 2 reports the effects on average 10th grade achievement of alternative rates of student mobility among the three schools. (The reported averages are population means, as opposed to means from a particular random sample. As is illustrated below, sample means are likely to be quite variable, particularly in small schools.) Panel B reports the associated standard deviations

[21]School performance is assumed to be equal to 10, 0, and -10, respectively, in the three types of schools. See Appendix B for additional details.

TABLE 2

**Average 10th Grade Achievement By School Type,
Given Alternative Rates of Student Mobility**

School Type	Annual Mobility Rate (Percent) 0	5	10	20	40	60
A. Average Tenth Grade Achievement						
High Performance	100*	72.2	53.5	32.4	16.7	11.1
Medium Performance	0*	0	0	0	0	0
Low Performance	-100*	-72.2	-53.5	-32.4	-16.7	11.1
B. Standard Deviation of Tenth Grade Achievement						
High or Low Performance	0	46.1	51.3	47.6	35.5	26.8
Medium Performance	0	32.3	39.0	40.9	33.9	26.6
All	81.6	72.4	64.6	52.6	37.5	28.3
C. The Fraction of Students Who Change Schools One or More Times (Percent)						
Change over 9 years (grades 1 to 10)	0	37.0	61.3	86.6	99.0	100.0
Change over 4 years (say, grades 4 to 8)	0	18.5	34.4	59.0	87.0	97.4

Source: Data simulated by author.

*In this simulation, average 10th grade achievement and steady state school performance are identical if the mobility rate is zero.

in student achievement. Panel C reports the fraction of students who change schools between grades 1 and 10 and over a four-year period, say between the 4th and 8th grade, given alternative annual rates of student mobility. Notice that student mobility causes average 10th grade test scores to collapse toward zero, the average level. For the high- and low-performance schools, for example, an annual mobility rate of 20 percent leads to a reduction in average test scores of over 70 percent. Similarly, an annual mobility rate of 40 percent caused a reduction in average test scores of 83 percent. In other words, the average test score is severely biased against high-performance 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.[22]

The numbers reported above (in Panel A of Table 2) actually understate the degree to which random mobility distorts the average test score as an indicator of school performance. In practice, average test scores are likely to vary substantially across small schools due to the random nature of mobility, even for schools with identical mobility rates. (Variation in means is a direct function of school size.) This problem is therefore apt to be especially serious among elementary schools and relatively small high schools. To illustrate the problem, Figure 4 reports the likely spread in average test scores for schools with twenty and fifty 10th grade students, respectively. For each type of school (high, medium, and low performance), the top average test score is equal to the population mean plus twice the standard error of the mean, and the bottom average test score is equal to the population mean minus twice the standard error of the mean.

The results reported in Figure 4 are striking. For high-performance schools with twenty 10th grade students the range in average test scores is from 30.6 to 76.4, given a mobility rate of only 10

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

Figure 4 here

percent. At a mobility rate of 40 percent, the distributions actually overlap. Given a 10th grade class of fifty students, average test scores range from 6.7 to 26.7 for high-performance schools, and from -9.6 to 9.6 for medium-performance schools. In summary, both of the above simulations demonstrate that random student mobility has a potentially enormous impact on average test scores.

Of course, rates and patterns of student mobility may vary over time as school systems merge, communities grow, and the occupational structure of jobs evolve in a local labor market. As is illustrated below, idiosyncratic patterns of mobility are also apt to provide a misleading picture of actual changes in school quality over time. Figure 5 simulates the effects on average 10th grade achievement of an influx of students from a low-quality to a high-quality school. Panel A of Figure 5 simulates the effects of a *gradual* influx of students that takes place over a ten-year period: 1992 through 2001. Panel B simulates the effects of an *instant* influx of students in 1992. Despite the fact that school performance 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.

The final simulation illustrates the potential consequences of differences across schools in student, family, and community characteristics for both the gain indicator and the average 10th grade test score. To highlight the consequences of this factor I make the following assumptions: (1) school performance differs among schools but does not vary across grades and years, (2) the effects (represented by the parameters β and γ) of student, family, and community characteristics on

(Figure 5 here)

achievement growth are identical in all grades, (3) there is no student mobility, and (4) initial average test scores are identical in all schools. Given these assumptions, the average 10th grade test score is given simply by

$$Y\ (10,\ t \mid s)\ =\ 10\{\beta \overline{X}(s)\ +\ \gamma C(s)\}\ +\ 10\alpha(s)$$

where the first term represents the average contribution of student, family, and community characteristics to student achievement growth in grades 1 through 10—the student and community adjustment factor—and 10 α(s) represents the steady state indicator of intrinsic school performance.[23]

I assume that steady state school performance ranges from -200 to 200 and the student and community adjustment factor ranges from -300 to 300. The greater spread of the latter factor reflects the implicit assumption that there is significant stratification across schools in terms of student, family, and community characteristics and that these characteristics account for substantially more of the variation in student achievement than in school performance.

As indicated in Table 3, for a medium performance school (second column from the right) the average 10th grade test score ranges from -300 to 300, solely because of differences across schools in average student, family, and community characteristics. In comparison, among schools that represent the middle of the distribution of student, family, and community characteristics (fourth row of the table), the average test score ranges from -200 (a low-performance school) to 200 (a high-performance school). Given the assumptions of this model, it is evident that differences across schools in student, family, and community characteristics could potentially obscure or eliminate differences that exist among schools in actual performance. To see this, note that there are three schools in Table 3 that are about average with respect to average 10th grade test scores (the data are

[23]The gain indicator is given simply by the average test score divided by 10.

TABLE 3

Average 10th Grade Test Score by School, Given Variation in Average Student, Family, and Community Characteristics and Steady State Intrinsic School Performance

Student and Community Adjustment Factor	Average Test Score by Steady State School Performance		
	Low (-200)	Med. (0)	High (200)
-300	-500	-300	-100
-200	-400	-200	⓪
-100	-300	-100	100
0	-200	⓪	200
100	-100	100	300
200	⓪	200	400
300	100	300	500

Source: Data simulated by author.

circled in the table, for convenience). In fact, these schools vary enormously in terms of their school performance, ranging from a low of -200 to a high of 200.

In short, it is clear that the average test score and the gain indicator are measures that, if interpreted as measures of school performance, are highly biased against schools that disproportionately serve academically disadvantaged students and communities.

The simulations presented in this section demonstrate that the average test score has the potential to provide a totally misleading portrait of educational productivity, both over time and across schools. Even so, the simulations possibly understate the degree to which this indicator is flawed as a valid measure of school performance since they address the problems of nonlocalization, aggregation, student mobility, and differences across schools in student, family, and community characteristics one at a time, not simultaneously.

The good news is that the gain indicator, if it can be computed, and the value-added indicator avoid three of the four problems that plague the average test score as a measure of school performance. The value-added indicator has the major advantage that it also eliminates the bias that exists in the gain indicator due to differences across schools in student, family, and community characteristics.

The next section presents a real-world example in which the gain indicator provides a substantially more accurate portrait of changes over time in school performance than the average test score.

IV. AN EXAMPLE BASED ON NATIONAL DATA

The real-world significance of the above analysis is illustrated using data on average mathematics scores from 1973 to 1986 from the National Assessment of Educational Progress (NAEP). As indicated in Panel A of Table 4, NAEP scores for the 11th grade exhibit the by now

TABLE 4

NAEP Mathematics Exam Data

Grade/Age	1973	1978	1982	1986
	Average Test Score (A)			
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
	Average Test Score Gain (B)			
	73 to 78	78 to 82	82 to 86	
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).

familiar pattern of sharp declines from 1973 to 1982 and then partial recovery between 1982 and 1986. The 11th 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.[24]

The gain indicators reveal that achievement growth during the 1982–1986 period was actually no better than achievement growth during the prior 1978–1982 period. In fact, gains from the 7th to the 11th grade were actually slightly lower during the 1982–1986 period than in previous periods! The rise in 11th grade math scores from 1982 to 1986 stems from an earlier increase in achievement growth for that cohort rather than from an increase in achievement growth over grades 7 to 11. 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.

It would be interesting to report the above analysis using value-added as opposed to gain indicators. Unfortunately, the NAEP data do not permit such an analysis to be conducted since the same students are not sampled for two consecutive NAEP surveys. This weakness in NAEP data could be remedied by switching to a survey design that was at least partially longitudinal.[25]

---

[24]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, we assume that average test scores in 1973 are comparable to the unknown 1974 scores.

[25]Given that the NAEP, at present, assesses students only in grades 4, 8, and 12, it would be necessary as part of a longitudinal survey design to locate students that changed schools within the four-year interval between tests. Of course, there are other important technical problems that would also need to be addressed.

V. THE CONSEQUENCES OF USING FLAWED INDICATORS

The fact that gain and level indicators measure school performance with potentially enormous error has important implications for the use of these indicators for making education policy, informing students and parents about the quality of alternative schools, and evaluating school performance as part of a high-stakes accountability system. With respect to the first issue, it is clear from the simulations and from the NAEP example that level indicators potentially provide totally incorrect information concerning the success or failure of educational interventions and reforms. As a result, making policy on the basis of such information could lead to the expansion of programs that do not work and to the cancellation of programs that are truly effective. Similarly, 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 simple 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 could severely harm the policy-making process and distort the school choices of students and families. As a result, student achievement is apt to lower than it would otherwise be.

The consequences of using average test scores or gain indicators for purposes of public accountability are, if anything, potentially much worse than in the cases discussed above. Why? Level and gain indicators could severely distort the behavior of teachers and administrators and lead to substantially reduced performance expectations for schools that disproportionately serve academically disadvantaged students. (This point is discussed more fully in the next section.) Moreover, these

indicators are biased against schools that disproportionately serve academically disadvantaged students and communities and thus are undesirable from the standpoint of fairness.

The first effect is likely to be particularly acute if teachers and administrators are in any way rewarded or penalized on the basis of their performance with respect to a given indicator. In a high-stakes accountability system—a system that rewards teachers and administrators for their performance—teachers and administrators are likely to respond to the incentive to improve their *measured* performance by exploiting all existing avenues to improve measured performance. It is well known, for example, that teachers may "teach *narrowly* to the test," although some tests are more susceptible to this type of corruption than others. For tests that are relatively immune to this type of corruption, teaching to the test could induce teachers and administrators to adopt new curriculums and teaching techniques much more rapidly than they otherwise would. On the other hand, if school performance is measured using level or gain indicators, teachers and administrators have the incentive to raise measured school performance by teaching only those students who rate highly in terms of average student and family characteristics, average prior achievement, and community characteristics. As indicated in section III, even modest changes in these characteristics could dramatically improve a school's gain indicator or average test score. This phenomenon is referred to 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 state-wide testing), and (6) make it difficult

for low-scoring students to participate in state-wide examinations. These activities are all designed to improve average test scores in a school, not by improving school quality, but rather by catering to high-scoring students while ignoring or alienating low-scoring students.

Instead of devoting excessive attention to 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.

VI. VALUE-ADDED INDICATORS: DATA REQUIREMENTS

Given the problems that exist with the average test score and other level indicators and, to a lesser degree, the gain indicator, it is important to consider whether value-added indicators could potentially be used as the core of school district, state, and national performance indicator/accountability systems. There are at least two reasons to be optimistic in this regard. First, value-added models have been used extensively over the last three decades by evaluators and other researchers interested in education and training programs. Second, a number of districts and states, including Dallas (Webster, Mendro, and Almaguer; 1992), South Carolina (Mandeville, 1994), and Tennessee (Sanders and Horn, 1994), have successfully implemented value-added indicator systems.

Nonetheless, despite the promise of value-added indicator systems, it is clear that they require a major commitment on the part of districts and states. In particular, districts and states must be prepared to, one, test students frequently, ideally at every grade level, as is done in South Carolina, Tennessee, and Dallas; and two, develop comprehensive district or state data systems that contain

information on student test scores and student, family, and community characteristics.[26] These two issues are discussed below.

Annual testing at each grade level is highly recommended for the following three reasons. First, it maximizes accountability by localizing school performance to the most natural unit of accountability, the grade level or classroom. Second, it yields up-to-date information on school performance. Finally, it limits the amount of data that is lost due to student mobility. As the time interval between tests increases, these problems become much more acute. In fact, for time intervals of more than two years it could prove difficult, if not impossible, to construct valid and reliable value-added (or gain) indicators for schools with high mobility rates. This problem arises because mobile students generally must be excluded from the data used to construct value-added and gain indicators, since both indicators require pre- and post-test data.[27] In schools with high student mobility, infrequent testing diminishes the likelihood of ending up with student data that are both representative of the school population as a whole and large enough to yield statistically reliable school performance estimates.[28] Less frequent testing, say testing at grades kindergarten, 4, 8, and 12, might be acceptable for national purposes, since student mobility is not really an issue at the national level.[29] For purposes of evaluating local school performance, however, the problems created by student

---

[26]The latter data are required as control variables in the value-added model.

[27]See Dyer, Linn, and Patton (1969) for a critique of school performance indicators that include mobile as well as nonmobile students.

[28]In schools with extremely high mobility rates it might be necessary to test students more than once a year.

[29]A kindergarten test is needed so that the growth in student achievement in grades 1 through 4 can be monitored. In my view, the National Assessment of Educational Progress and recent proposals for national testing in grades 4, 8, and 12 are seriously flawed by their failure to include a test at the kindergarten or 1st grade level.

mobility argue strongly for frequent testing, at least for schools and school districts where student mobility is high.

The primary obstacle to developing a comprehensive data system is, in my opinion, the difficulty of collecting extensive information on student and family characteristics. This issue is potentially quite important because value-added indicators are often implemented using the rather limited administrative data that are commonly available in schools, for example, race and ethnicity, gender, special education status, limited English proficiency (LEP) status, eligibility for free or reduced-price lunches, and whether a family receives welfare benefits. Researchers equipped with more extensive data have demonstrated that parental education and income, family attitudes toward education, and other variables are also powerful determinants of student achievement growth. It would be useful for school districts and states to experiment with some alternative approaches for collecting this type of data.

The consequence of failing to control adequately for these and other student, family, and community characteristics is that feasible real-world value-added indicators are apt to be biased because they absorb differences across schools in average *unmeasured* student, family, and community characteristics as well differences in intrinsic school performance. This implies that a feasible value-added indicator derived from a model with "weak" predictors of student achievement growth might be only slightly better than a gain indicator (better in the sense of being more highly correlated with a theoretically perfect value-added indicator). Even so, it is likely to be a much better indicator than the average test score.

VII. VALUE-ADDED INDICATORS AND SCHOOL PERFORMANCE EXPECTATIONS

Some commentators have raised the concern that value-added indicators, because they control (or adjust) for student, family, and community characteristics associated with student achievement

growth, inevitably result in reduced performance expectations for schools and states that disproportionately serve disadvantaged students.[30] Here, I demonstrate that the opposite is, in fact, true. The key idea is that it is possible to derive school performance goals, defined in value-added terms, that are fully consistent with high student-performance expectations. Moreover, these goals will always be higher, not lower, for schools that disproportionately serve disadvantaged students.

To see this, suppose that $G(g)^{EX}$ represents an achievement growth target for a given grade g. Growth targets could be set so that the cumulative growth in achievement across grades would enable all schools (including schools with very low initial achievement) to obtain a target level of achievement by the end of a given grade, say 6th grade.[31] Given equation (5), the school performance goal that is consistent with a specified growth target is equal to

$$\alpha\ (s,\ g,\ t)\ =\ G(g)^{EX}\ -\ Adjustment\ Factor\ (s,g,t). \quad (7)$$

The student and family adjustment factor (defined by (5)) represents the average contribution of student, family, and community characteristics to the average level of student achievement. Since this factor is lower for schools that disproportionately serve disadvantaged students, the effect of the adjustment is to generate higher school performance goals for these schools than for other schools. The reason is straightforward: in order to achieve a common student achievement goal it is necessary for schools that disproportionately serve disadvantaged students to out-perform other schools.

If student achievement expectations are sufficiently high, the above procedure will almost certainly produce school performance goals that are extremely ambitious for schools that disproportionately serve disadvantaged students. This is a strength, not a weakness, of the value-added

---

[30]See, for example, Finn (1994).

[31]This approach would allow schools with very low initial achievement a period of six years to catch up with schools with high initial achievement. The catch-up phase could, of course, be shortened or lengthened.

approach. If we as a society are serious about setting high expectations for all students, it is important to accurately translate these performance expectations into accurate school performance goals. Given concrete school performance goals we can then act accordingly to do whatever is necessary and appropriate to assist schools in attaining these goals.

VIII. 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.[32] 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 that tends to be grossly out of date, it is contaminated by student mobility, and it fails to distinguish the distinct value-added contribution of schools to growth in student achievement from the contribution of student, family, and community factors. As a result, the average test score is a weak, if not counterproductive, instrument of public accountability. The gain indicator, if it can be computed, and the feasible value-added indicator avoid three of the four problems that plague the average test score. The feasible value-added indicator has the major advantage that it potentially eliminates the bias that exists in the gain indicator due to differences across schools in student, family, and community characteristics, particularly if it is based on a model that includes an extensive set of control variables. In this case, it fully 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

[32]Other level indicators, such as the median test score, are similarly suspect.

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 validity 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, Tennessee, and Dallas, or at least at every other grade level, beginning with kindergarten. 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.

APPENDIX A
DESCRIPTIONS OF REPORTED SIMULATIONS

This appendix presents detailed specifications and results for the simulations presented in Figures 1, 2, 3, and 5. Appendix B considers the simulations reported in Figure 4 and Table 2. The first section presents the detailed specifications. Each simulation is defined in terms of intrinsic school performance in a given grade g at a given point in time t. Average 10th grade achievement is computed for each birth cohort. The birth cohort subscript is implied by the grade and time subscripts: c=t-g-6. For simplicity, we assume that students begin 1st grade at age 6 and advance to subsequent grades one year at a time. The second section reports intrinsic school performance and steady state school performance for grades 1 to 10 by grade and cohort and 10th grade achievement. School performance values at each grade level for a given year are reported on diagonal lines that run from the bottom left to the top right of the tables.

Specifications

Figure 1. Average 10th Grade Achievement Given Alternative Patterns of Intrinsic School Performance in Grades 1 to 6 and 7 to 10.

Panel A

Intrinsic School Performance, Grades 1 to 10:

$$\alpha(g,t) = \begin{cases} 0 & \text{if } t = 1978, 1991 \\ 0 + (t-1991) & \text{if } t = 1992, 2001 \end{cases}$$

Panel B

Intrinsic School Performance, Grades 1 to 6:

$$\alpha(g,t) = \{0 \text{ if } t = 1978, 2001\}$$

Intrinsic School Performance, Grades 7 to 10:

$$\alpha(g,t) = \begin{cases} 0 & \text{if } t = 1978, 1991 \\ 0+2(t-1991) & \text{if } t = 1992, 2001 \end{cases}$$

Figure 2. Average 10th Grade Achievement Given Alternative Patterns of Intrinsic School Performance Over Time.

Panel A

Intrinsic School Performance, Grades 1 to 10:

$$\alpha(g,t) = \begin{cases} 20 - (t-1971) & \text{if } t = 1971, 1991 \\ 0 + (t-1991) & \text{if } t = 1992, 2001 \end{cases}$$

Panel B

Intrinsic School Performance, Grades 1 to 10:

$$\alpha(g,t) = \begin{cases} 20 - (t-1971) & \text{if } t = 1971, 1986 \\ 5 + (t-1986) & \text{if } t = 1987, 1991 \\ 10 & \text{if } t = 1991, 2001 \end{cases}$$

Figure 3. Average 10th Grade Achievement Given Alternative Cycles of Decline and Recovery in Intrinsic School Performance

Panel A

School 1, Intrinsic School Performance, Grades 1 to 10:

$$\alpha(g,t) = \begin{cases} 20 - 2(t-1981) & \text{if } t = 1981, 1990 \\ 0 + 2(t-1991) & \text{if } t = 1991, 2000 \end{cases}$$

School 2, Intrinsic School Performance, Grades 1 to 10:

$$\alpha(g,t) = \begin{cases} 0 + 2(t-1981) & \text{if } t = 1981, 1990 \\ 20 - 2(t-1991) & \text{if } t = 1991, 2000 \end{cases}$$

Note: These patterns repeat in twenty-year cycles.

Figure 5. Average 10th Grade Achievement Given Different Patterns of Student Mobility

Panel A. A Gradual Influx of Students

This simulation considers the effects of a gradual influx of one hundred students from a low-gain school into a high-gain school that initially has one hundred students. Each year, from 1992 to 2001, ten students from the low-gain school, one in each grade, move to the high-gain school. The annual gains in the high- and low-gain schools are 20 and -20, respectively. As in Table 3, 10th grade achievement is computed as the sum of prior gains in grades 1 through 10.

Panel B.

This simulation is identical to the one reported for Panel A except that the influx of all one hundred students occurs at one point in time, 1992.

APPENDIX TABLE A-1

Data for Figure 1A

Year Cohort Average Completes Achievement Grade 10	Intrinsic School Performance by Grade										Steady State School Performance in Grades 1 to 10	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	0
1988	0	0	0	0	0	0	0	0	0	0	0	0
1989	0	0	0	0	0	0	0	0	0	0	0	0
1990	0	0	0	0	0	0	0	0	0	0	0	0
1991	0	0	0	0	0	0	0	0	0	0	0	0
1992	0	0	0	0	0	0	0	0	0	1	10	1
1993	0	0	0	0	0	0	0	0	1	2	20	3
1994	0	0	0	0	0	0	0	1	2	3	30	6
1995	0	0	0	0	0	0	1	2	3	4	40	10
1996	0	0	0	0	0	1	2	3	4	5	50	15
1997	0	0	0	0	1	2	3	4	5	6	60	21
1998	0	0	0	1	2	3	4	5	6	7	70	28
1999	0	0	1	2	3	4	5	6	7	8	80	36
2000	0	1	2	3	4	5	6	7	8	9	90	45
2001	1	2	3	4	5	6	7	8	9	10	100	55

Source: Data simulated by author.

APPENDIX TABLE A-2

Data for Figure 1B

Year Cohort Average Completes Achievement Grade 10 Grade 10	Intrinsic School Performance by Grade										Steady State School Performance in Grades 1 to 6	Steady State School Performance in Grades 7 to 10	in
	1	2	3	4	5	6	7	8	9	10			
1987	0	0	0	0	0	0	0	0	0	0	0	00	
1988	0	0	0	0	0	0	0	0	0	0	0	00	
1989	0	0	0	0	0	0	0	0	0	0	0	00	
1990	0	0	0	0	0	0	0	0	0	0	0	00	
1991	0	0	0	0	0	0	0	0	0	0	0	00	
1992	0	0	0	0	0	0	0	0	0	2	0	82	
1993	0	0	0	0	0	0	0	0	2	4	0	166	
1994	0	0	0	0	0	0	0	2	4	6	0	2412	
1995	0	0	0	0	0	0	2	4	6	8	0	3220	
1996	0	0	0	0	0	0	4	6	8	10	0	4028	
1997	0	0	0	0	0	0	6	8	10	12	0	4836	
1998	0	0	0	0	0	0	8	10	12	14	0	5644	
1999	0	0	0	0	0	0	10	12	14	16	0	6452	
2000	0	0	0	0	0	0	12	14	16	18	0	7260	
2001	0	0	0	0	0	0	14	16	18	20	0	8068	

Source: Data simulated by author.

APPENDIX TABLE A-3

Data for Figure 2A

Year Cohort Completes Achievement Grade 10 10	Intrinsic School Performance by Grade										Steady State School Performance in Grades 1 to 10	Average in Grade
	1	2	3	4	5	6	7	8	9	10		
1981	19	18	17	16	15	14	13	12	11	10	100	145
1982	18	17	16	15	14	13	12	11	10	9	90	135
1983	17	16	15	14	13	12	11	10	9	8	80	125
1984	16	15	14	13	12	11	10	9	8	7	70	115
1985	15	14	13	12	11	10	9	8	7	6	60	105
1986	14	13	12	11	10	9	8	7	6	5	50	95
1987	13	12	11	10	9	8	7	6	5	4	40	85
1988	12	11	10	9	8	7	6	5	4	3	30	75
1989	11	10	9	8	7	6	5	4	3	2	20	65
1990	10	9	8	7	6	5	4	3	2	1	10	55
1991	9	8	7	6	5	4	3	2	1	0	0	45
1992	8	7	6	5	4	3	2	1	0	1	10	37
1993	7	6	5	4	3	2	1	0	1	2	20	31
1994	6	5	4	3	2	1	0	1	2	3	30	27
1995	5	4	3	2	1	0	1	2	3	4	40	25
1996	4	3	2	1	0	1	2	3	4	5	50	25
1997	3	2	1	0	1	2	3	4	5	6	60	27
1998	2	1	0	1	2	3	4	5	6	7	70	31
1999	1	0	1	2	3	4	5	6	7	8	80	37
2000	0	1	2	3	4	5	6	7	8	9	90	45
2001	1	2	3	4	5	6	7	8	9	10	100	55

Source: Data simulated by author.

APPENDIX TABLE A-4

Data for Figure 2B

Year Cohort Average Completes Achievement Grade 10 Grade 10	Intrinsic School Performance by Grade										Steady School School Performance in Grades 1 to 10	in
	1	2	3	4	5	6	7	8	9	10		
1981	19	18	17	16	15	14	13	12	11	10	100	145
1982	18	17	16	15	14	13	12	11	10	9	90	135
1983	17	16	15	14	13	12	11	10	9	8	80	125
1984	16	15	14	13	12	11	10	9	8	7	70	115
1985	15	14	13	12	11	10	9	8	7	6	60	105
1986	14	13	12	11	10	9	8	7	6	5	70	95
1987	13	12	11	10	9	8	7	6	5	6	80	87
1988	12	11	10	9	8	7	6	5	6	7	90	81
1989	11	10	9	8	7	6	5	6	7	8	100	77
1990	10	9	8	7	6	5	6	7	8	9	100	75
1991	9	8	7	6	5	6	7	8	9	10	100	75
1992	8	7	6	5	6	7	8	9	10	10	100	76
1993	7	6	5	6	7	8	9	10	10	10	100	78
1994	6	5	6	7	8	9	10	10	10	10	100	81
1995	5	6	7	8	9	10	10	10	10	10	100	85
1996	6	7	8	9	10	10	10	10	10	10	100	90
1997	7	8	9	10	10	10	10	10	10	10	100	94
1998	8	9	10	10	10	10	10	10	10	10	100	97
1999	9	10	10	10	10	10	10	10	10	10	100	99
2000	10	10	10	10	10	10	10	10	10	10	100	100
2001	10	10	10	10	10	10	10	10	10	10	100	100

Source: Data simulated by author.

APPENDIX TABLE A-5

Data for Figure 3, School 1

Year Cohort Average Completes Achievement Grade 10 Grade 10	Intrinsic School Performance by Grade										Steady State School Performance in Grades 1 to 10	in
	1	2	3	4	5	6	7	8	9	10		
1981	2	4	6	8	10	12	14	16	18	20	200	110
1982	4	6	8	10	12	14	16	18	20	18	180	126
1983	6	8	10	12	14	16	18	20	18	16	160	138
1984	8	10	12	14	16	18	20	18	16	14	140	146
1985	10	12	14	16	18	20	18	16	14	12	120	150
1986	12	14	16	18	20	18	16	14	12	10	100	150
1987	14	16	18	20	18	16	14	12	10	8	80	146
1988	16	18	20	18	16	14	12	10	8	6	60	138
1989	18	20	18	16	14	12	10	8	6	4	40	126
1990	20	18	16	14	12	10	8	6	4	2	20	110
1991	18	16	14	12	10	8	6	4	2	0	0	90
1992	16	14	12	10	8	6	4	2	0	2	20	74
1993	14	12	10	8	6	4	2	0	2	4	40	62
1994	12	10	8	6	4	2	0	2	4	6	60	54
1995	10	8	6	4	2	0	2	4	6	8	80	50
1996	8	6	4	2	0	2	4	6	8	10	100	50
1997	6	4	2	0	2	4	6	8	10	12	120	54
1998	4	2	0	2	4	6	8	10	12	14	140	62
1999	2	0	2	4	6	8	10	12	14	16	160	74
2000	0	2	4	6	8	10	12	14	16	18	180	90
2001	2	4	6	8	10	10	14	16	18	20	200	110

Source: Data simulated by author.

APPENDIX TABLE A-6

Data for Figure 3, School 2

Year Cohort Average Completes Achievement Grade 10 Grade 10	Intrinsic School Performance by Grade										Steady State School Performance in Grades 1 to 10	in
	1	2	3	4	5	6	7	8	9	10		
1981	18	16	14	12	10	8	6	4	2	0	0	90
1982	16	14	12	10	8	6	4	2	0	2	20	74
1983	14	12	10	8	6	4	2	0	2	4	40	62
1984	12	10	8	6	4	2	0	2	4	6	60	54
1985	10	8	6	4	2	0	2	4	6	8	80	50
1986	8	6	4	2	0	2	4	6	8	10	100	50
1987	6	4	2	0	2	4	6	8	10	12	120	54
1988	4	2	0	2	4	6	8	10	12	14	140	62
1989	2	0	2	4	6	8	10	12	14	16	160	74
1990	0	2	4	6	8	10	12	14	16	18	180	90
1991	2	4	6	8	10	12	14	16	18	20	200	110
1992	4	6	8	10	12	14	16	18	20	18	180	126
1993	6	8	10	12	14	16	18	20	18	16	160	138
1994	8	10	12	14	16	18	20	18	16	14	140	146
1995	10	12	14	16	18	20	18	16	14	12	120	150
1996	12	14	16	18	20	18	16	14	12	10	100	150
1997	14	16	18	20	18	16	14	12	10	8	80	146
1998	16	18	20	18	16	14	12	10	8	6	60	138
1999	18	20	18	16	14	12	10	8	6	4	40	126
2000	20	18	16	14	12	10	8	6	4	2	20	110
2001	18	16	14	12	10	8	6	4	2	0	0	90

Source: Data simulated by author.

APPENDIX B
DESCRIPTION OF THE STUDENT MOBILITY SIMULATIONS

This appendix provides a technical description of the simulations reported in Table 2 and Figure 4. These simulations were designed to highlight the effect of student mobility on the average test score. The basic assumptions of the simulation model are as follows. First, intrinsic school performance in a given school is the same at all grade levels and over time. Second, there are three different levels of school performance, given by

$$\alpha(1) = 10$$
$$\alpha(2) = 0$$
$$\alpha(3) = -10.$$

Third, student mobility follows a Markov process. At the end of each school year students have a probability (1-p) of staying in the school they are currently enrolled in and a probability p/2 of moving to one of the other two schools. Hence, the annual student mobility rate is equal to p. Finally, student, family, and community characteristics are identical for all students. Hence, the student and community adjustment factor is identical for all schools and can be ignored.

Given these assumptions, students differ from each other only because they exhibit different patterns of enrollment in each of the three school types. A specific enrollment pattern for grades 1 through 10 can thus be characterized by a ten-digit sequence of numbers, where a "1" indicates that a student was enrolled in school type one, a "2" indicates that a student was enrolled in school type two, etc. For example, a student enrolled in school one in grades 1 to 4, in school two in grades 5 to 7, and in school three in grades 8 to 10 would have the following enrollment pattern: 1111 222 333. The total number of sequences is equal to

$$(\text{number of school types})^{(\text{number of grades})} = 3^{10} = 59{,}049.$$

Given the student mobility rate p, the probability of each enrollment pattern i is given by

$$w(i) = \left(\frac{1}{3}\right)\left(\frac{p}{2}\right)^{c(i)}(1-p)^{(9-c(i))}$$

where c(i) = the number of school enrollment changes in a given enrollment pattern i. For example, c(i) = 2 for the enrollment sequence 1111 222 333. Given a mobility rate of p = .02, the probability of this sequence is equal to

$$\frac{1}{3}(0.2)^2(1-0.2)^7 = 0.002796.$$

Student achievement at the end of 10th grade is determined simply by the number of years enrolled in each type of school. Let $\tau(1,i)$, $\tau(2,i)$, and $\tau(d,i)$ represent the number of years enrolled in each of the three school types for students with enrollment pattern i. Then, 10th grade achievement for students in enrollment pattern i is given by

$$y(i,10) = \tau(1,i)\ \alpha(1) + \tau(2,i)\ \alpha(2) + \tau(3,i)\ \alpha(3).$$

It is straightforward to compute the population mean and standard deviation of 10th grade achievement, given a student's enrollment status as of 10th grade. These numbers are reported in Panels A and B of Table 2. First, generate a data set that includes the following variables for all enrollment patterns, given an assumed student mobility rate:

Years enrolled in school type one: $\tau(1, i)$
Years enrolled in school type two: $\tau(2, i)$
10th grade student achievement: $\tau(3, i)$
Number of school enrollment changes: c(i)
Probability of enrollment pattern: w(i).

Second, use a standard statistics program to compute the weighted mean and standard deviation of 10th grade achievement, given 10th grade enrollment status. The enrollment pattern probability w(i) is the appropriate weight.

Panels A and B of Table 2 report the population mean and standard deviation of 10th grade achievement, by 10th grade enrollment status and the assumed student mobility rate. The standard error of a sample mean is given by the standard deviation divided by the square root of the sample size. This formula was used to compute the standard errors required to construct the 2/-2 standard error interval illustrated in Figure 4.

Finally, Panel C of Table 2 reports the fraction of students who change schools one or more times over intervals of four and nine years, respectively. These numbers were computed using the following formula:

Prob (change over x years) = $1 - (1 - p)^x$.

For example, given a student mobility rate of 0.2, the probability of changing schools at least once over nine years is given by

$1 - (1 - 0.2)^9 = 0.866$,

as indicated in Table 2.

References

Boardman, Anthony E. and Richard J. Murnane, 1979. "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement," *Sociology of Education*, Vol. 52, pp. 113–121.

Bryk, Anthony S. and Stephen W. Raudenbush, 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*, Newbury Park, Calif.: Sage Publications.

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

Darling-Hammond, Linda, 1991. "The Implications of Testing Policy for Quality and Equality," *Phi Delta Kappan*, Vol. 73, No. 3 (November), pp. 220–225.

Dossey, John A. et al., 1988. *The Mathematics Report Card: Are We Measuring Up*? Princeton: 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*, Vol. 6, No. 4 (November), pp. 591–605.

Finn, Jr., Chester E., 1994. "Drowning in Lake Wobegon," *Education Week* (June 15), pp. 31, 35.

Gamoran, Adam, 1992. "Social Factors in Education," in Marvin Alkins (ed.), *Encyclopedia of Educational Research* (sixth edition), New York: Macmillan, pp. 1222–1229.

Goldman, Jay P., 1990. "Grading Schools through Report Cards: Realtors, News Media Collect Them for Comparative Purposes," July 23, 1990.

Haladyna, Thomas M., Susan Bobbit Nolen, and Nancy S. Haas, 1991. "Rising Standardized Achievement Test Scores and the Origins of Test Score Pollution," *Educational Researcher*, Vol. 20, No. 5 (June–July), pp. 2–7.

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

Hanushek, Eric A. and Lori Taylor, 1990. "Alternative Assessments of the Performance of Schools," *Journal of Human Resources*, Vol. 25, No. 2 (Spring), pp. 179–201.

Koretz, Daniel et al., 1994. "The Vermont Portfolio Assessment Program," *Educational Measurement: Issues and Practice*, Vol. 13, No. 3 (Fall), pp. 5–16.

Mandeville, Garrett K., 1994. "The South Carolina Experience with Incentives," paper presented at the conference entitled Midwest Approaches to School Reform, Federal Reserve Bank of Chicago (October 26).

Meyer, Robert H., 1992. "Applied versus Traditional Mathematics: New Econometric Models of the Contribution of High School Courses to Mathematics Proficiency," Institute for Research on Poverty, Discussion Paper No. 966-92, University of Wisconsin–Madison.

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

Nolen, Susan Bobbit, Thomas M. Haladyna, and Nancy S. Haas, 1992. "Uses and Abuses of Achievement Test Scores," *Educational Measurement: Issues and Practice*, Vol. 11, No. 2 (Summer), pp. 9–15.

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

Raudenbush, Stephen W. and Anthony S. Bryk, 1986. "A Hierarchical Model for Studying School Effects," *Sociology of Education*, Vol. 59, pp. 1–17.

Sanders, William L. and Sandra P. Horn, 1994. "The Tennessee Value-Added Assessment System (TVAAS)," forthcoming in *Journal of Personnel Evaluation in Education*.

Shepard, Lorrie A., 1991. "Will National Tests Improve Student Learning?" *Phi Delta Kappan*, Vol. 73, No. 3 (November), pp. 232–238.

Smith, Marshall S. and Jennifer O'Day, 1990. "Systemic School Reform," in Susan Fuhrman and Betty Malen (eds.), *The Politics of Curriculum and Testing (1990 Yearbook of the Politics and Education Association)*, London: Taylor and Francis, pp. 233–267.

Smith, Mary Lee and Clair Rottenberg, 1991. "Unintended Consequences of External Testing in Elementary Schools," *Educational Measurement: Issues and Practice*, Vol. 10, No. 4 (Winter), pp. 7–11.

Wainer, Howard, 1986. "The SAT as a Social Indicator: A Pretty Bad Idea," in Howard Wainer (ed.), *Drawing Inferences from Self-Selected Samples*, New York: Springer-Verlag, pp. 7–22.

Webster, William J., Robert L. Mendro, and Ted O. Almaguer, 1992. "Measuring the Effects of Schooling: Expanded School Effectiveness Indicies," paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wiggins, Grant, 1989. "A True Test: Toward More Authentic and Equitable Assessment," *Phi Delta Kappan*, Vol, 70, pp. 703–713.

Willett, John B., 1988. "Questions and Answers in the Measurement of Change," in Ernst Z. Rothkopf (ed.), *Review of Research in Education 15*, Washington, D.C.: American Educational Research Association, pp. 345–422.

Willms, Douglas J. and Stephen W. Raudenbush, 1989. "A Longitudinal Hierarchical Linear Model for Estimating School Effects and Their Stability, *Journal of Educational Measurement*, Vol. 26, No. 3 (Fall), pp. 209–232.