

Working PAPER

BY STEVEN M. GLAZERMAN AND LIZ POTAMITES

False Performance Gains: A Critique of Successive Cohort Indicators

December 2011

Abstract

There are many ways to use student test scores to evaluate schools. This paper defines and examines different estimators, including regression-based value-added indicators, average gains, and successive cohort differences in achievement levels. Given that regression-based indicators are theoretically preferred but not always feasible, we consider whether simpler alternatives provide acceptable approximations. We argue that average gain indicators potentially can provide useful information, but differences across successive cohorts, such as grade trends, which are commonly cited in the popular press and used in the Safe Harbor provision of federal school accountability laws, are flawed and can be misleading when used for school accountability or program evaluation.

Introduction

The No Child Left Behind Act of 2001 (NCLB) dramatically increased the amount of testing in American schools. Nevertheless, many educators and policy makers remain naïve about how to take full advantage of the information provided by such frequent testing. Also, public opinion, as shaped by the popular press, is often influenced by measures that are easy to calculate but possibly misleading.

Researchers have described in detail the rationale and mechanics of a high-quality system of indicators that could be used to measure school performance (Meyer 1994, 1996, and 1997; Ladd 1996; Kane and Staiger 1999; Stone 1999; Ladd and Walsh 2002; McCaffrey et al. 2004; Glazerman et al. 2010). For example, Meyer defines classes of performance measures as “level indicators,” such as proficiency rates and average test scores; “gain indicators,” which use prior test score information; and “value-added indicators,” which also use prior test score information, but do so flexibly and adjust for other student differences in measuring teacher contribution to achievement growth patterns.

These researchers argue that the key feature of a valid measure of educational effectiveness, one that provides school staff with the right incentives, is that it should isolate the unique contribution of the school to student achievement, *given the school staff’s available resources and the challenges they face*. This is the school’s “value-added;” it isolates the effect of school-based inputs as distinct from non-schooling inputs such as the influence of parents and student background.

Yet there remains widespread use of simpler measures that confound non-school inputs with schooling inputs, sometimes with high stakes involved. One of the most commonly used indicators in the popular press takes the differences between two different cohorts of students as a measure of the success of a school or a policy. We call these “successive cohort difference indicators.” An example would be comparing the average test scores for this year’s tenth grade students to the scores for last year’s tenth grade students. Headlines such as “On Reading Test, Mixed Results Under Bloomberg” (*New York Times*, 5/20/10) are based on cohort differences, in this case, comparing the performance of different cohorts of 4th and 8th graders. As we discuss below, non-school inputs that change the composition of the student cohorts being tested will affect the “results” in question.

The fact that the changing composition of cohorts may be contributing to the differences is sometimes even acknowledged in the same article. For example, in an article with the headline “Reading and math scores fall sharply at two KIPP schools in District”

(*Washington Post*, August 7, 2010), the executive director at the Knowledge is Power Program (KIPP) schools in Washington, DC pointed out that the decline was among fifth graders, who are the newest to the KIPP system, and said that they were observing more closely the changes in their incoming students.¹

The Safe Harbor provision of NCLB also relies on cohort differences. If a school fails to achieve Adequate Yearly Progress (AYP) due to the level of all students or those in a specific subgroup, it can still avoid sanctions if the percentage below proficiency has decreased by 10 percent from the previous year. This calculation ignores the fact that the population of students may have changed. For example, in states where only 10th graders are tested at the high school level, the Safe Harbor provision compares two different cohorts of students. A school thus may find a “safe harbor” just because the composition of its 10th graders has improved, not because the teachers have improved their instruction.

This problem is not limited to high schools. It affects any school where the population of students being tested from year to year changes, either overall or in a specific subgroup. For example, a school that may have made tremendous progress with its English language learners (ELL) in a previous year can still fail to find a “safe harbor” if it has a new group of ELL students currently failing to reach proficiency or if its policy changes in the direction of less aggressively labeling students as ELL.

To address this popular practice of using changing test scores across different cohorts as a measure of educational progress or failure, we provide a formal framework in this paper to assess this and other commonly used measures. We ask whether any of them might serve as reasonable approximations of value-added performance for purposes of evaluation or accountability. In the next section, we describe some of the indicators in common use in more detail. We then lay out a formal model to illustrate an ideal measure of school performance for policymaking. We apply this model to discuss three classes of indicators that aim to estimate true school performance and derives the bias of each measure. We use data from a large urban school district to compare the three measures of school performance. In the final section, we make recommendations for policy and further research.

Common Uses and Misuses of Test Score-Based Indicators

Measuring value-added normally requires a great deal of expensive data collection, including annual testing in every grade, tracking of data on student and family background, and careful tracking of student enrollment, mobility, and, in some cases, teacher-student links. Currently, it is common to test math and reading in grades 3 through 8 and at least one grade in high school, which provides an initial (baseline) and follow-up test for grades 4 through 8 and typically covers about 20 percent of a system’s teachers. Costs can rise dramatically as additional grades and subjects are included.

Frequent achievement testing is particularly burdensome because, educators argue, it takes time away from learning. Opponents of standardized testing find traditional achievement tests inappropriate or unfair (Medina and Neil 1990; Kohn 2000). Another challenge for implementing value-added measurement is the need for transparent methods to win over stakeholders and succeed politically (Ladd 1996). The statistical issues—such as regression adjustment, measurement error correction, and estimation error—make this difficult.

As a result, state and district education officials and the popular media have purposely or unwittingly adopted a range of approximations that are easier to understand, require fewer data, and provoke less controversy than value-added measures. Two widely used indicators are (1) average gains in student achievement or proficiency rates for a given group of students (“same-cohort change” or “average gain”) and (2) changes in average scores or proficiency rates in the same grade from year to year (“successive cohort change” or “cohort difference”). Average gain indicators are attractive because they are simple to understand and do not involve statistical controls. Cohort difference indicators are attractive because they can be used when testing is done at only a few selected milestone grade levels, such as 4, 8, and 12, and for grades such as grade 3 or subjects such as high school science that might include an exam at the end of the year but no baseline achievement measure to use for measuring growth.

Table 1 illustrates the difference between average gain and successive cohort difference indicators for a fictional school. Panels A and B each list the same hypothetical values of the average test scores by grade and year, but the performance measures are different. Gain measures, based on comparisons of snapshots from the same cohort in two successive grades, indicate that test scores rose in both grades 7 and 8, with an average increase of 5 and 10 points respectively. The ovals indicate the pairs of scores being compared. The successive cohort difference measures, shown in Panel B, lead to a very different conclusion: scores fell in grades 6 and 8 and were flat in grade 7. Overall, one might conclude that the school lowered test scores.

These two types of indicators were fairly common in education policy in the years before the federal education law (the Elementary and Secondary Education Act) was reauthorized as NCLB. State departments of education often used average gain and cohort difference indicators, sometimes with high stakes involved for school staff. For example, Pennsylvania used differences in test scores in the same grade from one cohort to the next (successive cohorts) to grant monetary awards to schools. Minnesota relied on average achievement gains to hold charter schools accountable (see Table 2).

Advocates, program evaluators, and other researchers also routinely used average gain and cohort difference indicators to document the performance of schools and school reforms. Some examples can be found in the evaluation research used to show evidence of success for several whole-school reform models. For example, the Talent Development High Schools, High Schools That Work, and the Literacy Collaborative all documented success using cohort difference indicators (Education Commission of the States 2001), as did the Benwood Initiative that reformed low-performing schools in Chattanooga, Tennessee (Silva 2008). A report on the Benwood Initiative claimed that the participating schools “posted significant gains” as a result of changes in the proficiency rates of different cohorts.

More recently, we find high-stakes use of successive cohort indicators in places like the state of California, which uses “growth” in its Academic Performance Index (API) as a measure of school success. The API itself is a weighted average of percentages of students meeting each of the state’s proficiency cutoffs on the California Standards Test (CST). Thus, it is essentially a level indicator of performance based on the proficiency levels of a current set of students. “API growth” is defined as the change from one year to the next in API, which is the difference in weighted proficiency levels between the current and the prior year’s cohorts. Another example is the Safe Harbor provision of NCLB accountability, discussed above. Cohort-to-cohort comparisons continue to be used in evaluations. The National Research Council used

trends in grade-level achievement (cohort difference measures) in its recent report on its evaluation of school reform in the District of Columbia (National Research Council 2011).

These examples demonstrate how policy makers, program operators, and even evaluation experts are willing to go against the advice of researchers and use indicators with potentially serious flaws, as we discuss below. This paper asks which, if any, of these practical compromises is acceptable.

True School Performance

Before considering alternative school performance indicators, it is helpful to begin with an ideal standard for what the indicators are trying to measure. This section presents a formal model that can describe school performance and places it in the context of potentially confounding factors that also affect student achievement. The discussion refers to *school* performance, but the ideas apply generally to the performance of types of schools (such as traditional public versus charter schools), teachers, districts, or any education intervention.

To focus the discussion, we assume that policy makers want to measure the performance of a school in producing student achievement at each grade level and localize that estimate to the most recent year to learn how that grade currently is performing. To be more compact and precise, the discussion about performance can be recast in terms of a more formal achievement growth model similar to those proposed by Willett (1988) and Bryk and Raudenbush (1992), shown in equation (1).

Let Y_{ijg} represent student achievement level, measured by a test score or performance assessment,² for student i in school j and grade g . Thus, $Y_{ij,g-1}$ represents the pretest results for student i in school j , where testing is done at the end of the school year in at least two consecutive grades. Let X_{ijg} represent a vector of student and family background characteristics that affect learning, and I_{ij} indicate whether (or how long) student i attends school j .³ Thus, a basic individual student achievement growth model can be written as follows:

$$(1) Y_{ijg} = \theta Y_{ijg-1} + \beta' X_{ijg} + \alpha_{jg} I_{ij} + e_{ij}$$

The coefficients on the variables for pretest, student background, and the school indicator— θ , β , and α_{jg} , respectively—are the parameters to estimate. The main parameter of interest, α_{jg} , measures the school's contribution to achievement in grade g . Any unobserved explanatory factors, including idiosyncratic or random variation, are denoted by the last term e_{ij} , which is assumed to be unknown, but with a known distribution and uncorrelated with X_{ijg} and Y_{ijg-1} . Education researchers have long cautioned that the variance of e_{ij} could be large if student and family background characteristics are not properly accounted for (Coleman et al. 1966). The vector X_{ijg} should therefore include all of the important individual-level determinants of achievement, or at least all of those that are correlated with I_{ij} . In practice, it may not be possible to specify all such factors—a problem previously considered.

Test-Score-Based Indicators of School Performance

This section presents a list of three estimators and examines their advantages and disadvantages. This list, summarized in Table 3, serves as a menu of alternatives for those wishing to construct school performance indicators. The estimators are listed in descending order, from the most theoretically valid to the more practical in terms of data requirements and analytic simplicity.

Value-Added Indicator

The method for constructing a *school* performance indicator using *student* test score data most often involves decomposing variance in achievement growth over a given time period into separate components attributable to student (and family) background and the school. The problem lends itself readily to conventional regression or analysis of variance methods using a model like that presented in equation (1).⁴ These approaches seek to measure value-added by statistically controlling for factors outside of the direct influence of the teachers or programs being evaluated. Further discussions of the theory and estimation of value-added indicators can be found elsewhere (Raudenbush and Bryk 1989; Meyer 1994 and 1996; Ladd 1996).

Estimating equation (1) directly would be the most straightforward method of isolating α . Thus, the estimated regression coefficient $\hat{\alpha}_{jg}$ would be the value-added estimate. The critical assumption is that any factors omitted from X_{ijg} (and thus in the error term e_{ij}) are not correlated with the school indicators, I_{ij} . If, for example, some schools or school types have more motivated students or parents (to the extent that generating such motivation is not the responsibility of the school itself), and motivation is not properly accounted for in X_{ijg} , then $\hat{\alpha}_{jg}$ may overstate the effects of such schools relative to the others. To the extent that the model captures all the important X variables and complete data are available for all students, this regression coefficient would be an unbiased “value-added” estimator of school performance. The uncertainty around the estimate could be gauged using the usual regression framework. The standard error, which describes that uncertainty, depends most critically on the residual variance in test scores (after conditioning on prior test scores and other determinants of achievement) and the sample size—in this case, the number of students and schools.

Unfortunately, students are not sorted randomly across schools, so we must account for as wide a range of explanatory factors as possible, such that we can rule out substantial bias due to omitted variables and selection on unobservable variables. The most critical variable is prior achievement. The other measures most readily available do not capture all of the factors that really matter, but tend to be crude proxies, such as free and reduced price lunch eligibility or race/ethnicity, as well as disability/special education status, English language learner status, and possibly an indicator for being over age for grade.

Average Gain Indicator/Same-Cohort Change

A shortcut can compute the gain in achievement from one year to the next by each school’s students in a given cohort. A school can be thought of as all of the tested grades, or just a single grade. The logic is the same. This is an intuitive and simple way to measure achievement growth in practice. This school performance indicator, the school’s average gain for cohort A ($\hat{\alpha}_{AA}$), is the average of the post-test minus the average of the pretest for the same cohort in the prior year/prior grade:

$$(2) \hat{\alpha}_{AA} \equiv \bar{Y}_{A,g} - \bar{Y}_{A,g-1} = \text{Same Cohort Change}$$

This indicator requires a pretest for grade ($g-1$) as well as the post-test for grade g , so it still requires testing in adjacent grades. The key advantage of this indicator is that it is simple for most stakeholders to understand. There is no statistical adjustment or regression to explain, and no need to collect student background or school and community context data.

The key disadvantage of this indicator is that it assumes that differences between schools in the average family background characteristics and other contextual factors do not affect the growth in achievement, just the level. This assumption can never hold perfectly, so we consider below a framework for characterizing the bias of the gain indicator.

A variant of the average gain indicator would use test score data from nonsuccessive grades. For example, if a district wanted to save resources by testing at the end of the year for grades 8 and 10 to evaluate high schools in science, the indicator would be $\hat{\alpha} = (\bar{Y}_{A,10} - \bar{Y}_{A,8})$. This indicator is intuitively appealing because it appears to save resources by reducing testing frequency while still holding two grade levels accountable—in this case, grades 9 and 10.

Unfortunately, this two-year gain indicator measures only half of the school's performance and cannot localize it to a specific year. The indicator measured at the end of year t includes performance of the grade 10 teachers in year t and that of the grade 9 teachers in year $(t-1)$, providing no information about grade 9 in year t or grade 10 in year $(t-1)$. Repeating the process every year does not help. The same two-year gain indicator constructed at the end of year $(t+1)$ would include information on performance of grade 9 in year t and grade 10 in year $(t+1)$. Again, this is a partial, nonspecific indicator of the two grades' performance. Testing at grade levels more than two years apart would just make the problem worse. That is why the National Assessment of Educational Progress (NAEP), which measures performance in the fourth, eighth, and twelfth grades, cannot be used for meaningful accountability or evaluation.

Bias of the Average Gain Indicator

While it is simpler and less costly to avoid the need for statistical controls, the average gain indicator, even if it is based on successive year data, is in a danger of being biased. It is equivalent to estimating equation (1) with the restrictions that $\theta = 1$ and $\beta = 0$. If either of these restrictions does not hold, the estimator is a biased measure of the school impact. Expressed in terms of the notation used here, the bias term is written:

$$(3) \text{Bias}(\hat{\alpha}_{AA}) = E[\hat{\alpha}_{AA} - \alpha] = (1 - \theta)E[\bar{Y}_{A,g-1}] + \beta' \bar{X}_{A,g}$$

Thus, the bias consists of two parts. The first part represents the amount of prior accumulated knowledge not properly accounted for in the model, a bias that results from assuming that the pretest coefficient (θ) is one. Depending on the scaling of the achievement tests used in the two grade levels, the assumption may be plausible—in other words, the constraint may be nonbinding, so that the first bias term is zero. In particular, if the same test is given in both grades and there is no summer learning loss, it would be reasonable to expect that the learning from the prior year is carried over approximately on a one-for-one basis. On the other hand, if the units of the test scores are different, the unadjusted gain score would be problematic. To reduce or eliminate the bias, one can substitute a scaling parameter that represents the equivalent score on the higher grade test of a lower grade test score. This would be possible if, for example, the test publisher provides a psychometric report that equates the two tests.

It also is common to express the test scores in terms of percentiles or standard deviation units based on a national norm. This places the pretest and post-test in similar-appearing units and makes the model restriction somewhat more plausible, although it can raise additional problems. For example, percentile scores are ordinal measures, so sums and differences of percentiles can be misleading statistics.

The second part of the bias comes from assuming that the effect of student and family background on achievement growth is zero. This is far less plausible and would bias the indicator in favor of schools or interventions that serve students of a higher socioeconomic status. A great deal of research evidence dating back to Coleman (1966) demonstrates the importance of student and family background characteristics on student achievement growth, as well as level. Thus, unadjusted average gain scores may erroneously attribute slow achievement growth in schools with poor or disadvantaged students to the teachers and the policies affecting those students, when those same students actually might have fared worse in other, higher-ranked schools.

Of those who use the same-cohort difference indicator in practice, many are aware of this second source of bias and take steps to eliminate it. One way is to group schools by similarity in socioeconomic status (variables in the X matrix in Equation [1]). An example of this is the use of matching procedures, in which schools are compared within categories defined by the fraction of students eligible for free or reduced-price lunches. This relies on *ad hoc* judgments about what matters and how much. In other words, policy makers are substituting their best guesses for the β vector.

Precision of the Average Gain Indicator

Two important statistical properties of a proposed estimator are its bias and precision. We have argued that the same-cohort change (gain) indicator very likely contains omitted variable bias. The precision of this estimator, however, is likely to be similar or greater than that of the value-added indicator estimated from the full regression model. This is a well-known result that follows from the algebra of a regression with an omitted variable (Greene 1993). The increase in precision depends on the correlation between the school type indicator variables and the omitted (mostly student and family background) variables. Regardless, the added precision is not much of an advantage for this estimator because, without knowing the coefficient of these omitted variables, the estimate of the precision will itself be biased. Therefore, another drawback of this type of school performance indicator is that researchers are unable to provide a reliable margin of error. This applies to performance gains estimated by random assignment of school type as well. The only way to capture the precision advantage of the average gain indicator is to use out-of-sample information to justify any restrictions placed on β and θ . This could make the average gain indicator a powerful tool but could create new data demands and add complexity, thereby undermining its key advantages.

Cohort Difference Indicator

The same-cohort gain indicator just described requires testing in every grade. In many situations, even this requirement is too onerous. For example, science, history, and foreign languages might be tested only in one grade. At the early elementary level, nearly every testing regime will have an entry grade for which there is no pretest. Thus, the tempting alternative would be to test in a given grade and subject each year and track differences at that grade level from year to year. This effectively compares different cohorts of students, using the “successive cohort difference” measure described in the introduction. Labeling the cohorts A and B, where group A starts in grade $(g-1)$ and group B starts in grade g , the successive cohort difference indicator ($\hat{\alpha}_{AB}$) would be written:

$$(4) \hat{\alpha}_{AB} \equiv \bar{Y}_{A,g} - \bar{Y}_{B,g} = \text{Successive Cohort Change}$$

Bias of the Cohort Difference Indicator

To many, the successive cohort difference indicator might seem like a reasonable approximation of the school impact but unfortunately it is even more severely biased than average gain. To simplify the notation, consider the indicator calculated using the current year's fifth graders and the prior year's fifth graders, so $\bar{Y}_{A,5}$ is the average fifth-grade achievement score for cohort A. The prior year's fifth graders are labeled cohort B. Through simple substitution, the bias is:

$$(5) \text{Bias} = E \left[(\bar{Y}_{A,5} - \bar{Y}_{B,5}) - \alpha_{A,5} \right] = \theta E \left[\bar{Y}_{A,4} - \bar{Y}_{B,4} \right] + \beta' (\bar{X}_{A,5} - \bar{X}_{B,5}) - \alpha_{B,5}$$

The bias has three components. The first bias term represents the accumulated differences in prior achievement of the two cohorts before beginning the fifth grade. The second term represents the effect of differences in student and family background between the two cohorts. The last term represents the effectiveness of the fifth grade for cohort B, which is only the effectiveness of the fifth grade in the year prior to the year that the indicator aims to capture.

To understand why the bias is not likely to be zero in any nonexperimental setting, it helps to examine equation (5) from another perspective. Let the variable prefix Δ represent the difference in the mean value of that variable between cohort A in a given grade and year and cohort B in the same grade and year. By repeated substitution of the growth model from equation (1), the bias in the successive cohort difference can be expressed as:

$$(6) E \left[\hat{\alpha}_{AB} - \alpha_{A,5} \right] = \theta^5 \Delta \bar{Y}_0 + \sum_{g=1}^5 \theta^{5-g} (\beta' \Delta \bar{X}_g) + \sum_{g=1}^4 \theta^{5-g} (\Delta \alpha_g) - \alpha_{B,5}$$

Here the three sources of bias are expressed differently. The first term represents the initial differences at school entry, appreciated or depreciated over the years, depending on the true growth parameter θ . The influence of θ here is to weight the amount of learning that takes place in each year. Here we assume for simplicity that θ is the same in every year. It is quite likely, however, that learning in the early years is especially important, so θ might vary over time and be greater than 1 in the early grades, further inflating this bias term. Even if $\theta = 1$ in every year, initial differences carry through. An example of a situation in which this bias could present a problem is the introduction of a full-day kindergarten or prekindergarten program in a community. If this were a good policy with lasting positive effects, it might be erroneously recognized only five years later as an impact attributable to whoever happens to be teaching fifth grade at that time.

The second term represents the accumulated differences in cohort characteristics (appreciated or depreciated over the years by θ). If θ were equal to 1, then the fifth-grade indicator would be contaminated by a student effect five times as large as the one in the same-cohort indicator. If θ were different, the bias term might be higher or lower, but most likely higher, for the reasons given above. The cohort characteristics bias would render an accountability system unfair to any school experiencing demographic shift toward students from lower socioeconomic-status households. For example, a factory opening in one year and attracting low-wage workers and their children could have the unintended effect of making the schools appear to be in decline several years later, even if the schools perform admirably to help newer students catch up to their more affluent predecessors.

The third component of the bias is a bit more complicated. It represents the difference between the two cohorts in the quality of schooling they received as they progressed through the

grade levels minus the effectiveness of the fifth grade for cohort B, which is just the effectiveness of the fifth grade in the year prior to the year that the indicator aims to capture. In other words, it is the accumulation of historical differences between the value-added in the classrooms of the two cohorts over the years (depreciated by a factor of θ each year) with an extra penalty if the comparison cohort was above average in the year immediately before the one that we intend to measure (or unfair boost if it was below average).

The formulation may not be intuitive, but it can be expressed in terms of common ideas, as shown in equation (7):

$$(7) \text{ Bias} = \frac{\text{Prior Achievement Differences at Entry}}{\text{Differences}} + \frac{\text{Differences in Student Effects Since Entry}}{\text{Differences}} + \frac{\text{Differences in School Effects Since Entry}}{\text{Differences}}$$

The cohort difference indicator includes a superfluous collection of historical differences in ability, background, and schooling experiences between two different groups of students. Therefore the successive cohort indicator can be a severely biased estimate of true school performance, α . It includes some magnified bias terms related to the between-cohort differences in students' family background and the accumulated school effects of prior years and grades since the cohorts entered formal schooling.

The differences in school effects since entry are a potentially troublesome source of bias for the cohort difference indicator. Assuming that these historical differences are zero would not only be implausible but in most cases would be illogical, because it violates the assumption that school effects vary over time—the assumption that justifies estimating annual school performance each year in the first place. Otherwise, policy makers would need only to estimate school effects once, and that estimate would represent the school's effectiveness indefinitely. Therefore, the assumption that school effects vary over time is self-evident.

There are more reasons why the assumption of time-varying school effects is more plausible than constant school effects. Even if the same teachers are in the same grades in every year, their individual teaching effectiveness may change from year to year. Over time they accumulate experience with the curriculum, the school, and the students that may affect their performance. This idea that experience is related to performance undergirds the whole structure of teacher compensation in American public education. Even so, teachers do not remain fixed in their classroom assignments, in general. Therefore, natural staff turnover would be another reason why students in the prior cohort would not have been exposed to the same educational interventions as students in the current cohort, and therefore would serve as a poor comparison group.

Some have suggested using cohort differences with multiple time points, a model outlined by Cook and Campbell (1979), to identify school impacts. Bloom (1999) gives one example referred to as a short interrupted time series. The primary idea is to measure four or five cohorts' test scores prior to the introduction of an intervention and then measure program impacts as deviations in each successive year from that preprogram time trend. This is a more sophisticated type of cohort difference indicator because it does not assume that every cohort is identical. It does assume, however, that there is a stable parameter that describes some constant change in cohorts from year to year. The assumptions needed for the short interrupted time series estimator to be unbiased are not as strong as those stated above, but still quite strong. The structural model is essentially:

$$(8) \bar{Y}_c = \rho \bar{Y}_{c-1} + \alpha I_c + e_c,$$

where c indexes cohorts. This is very similar to the model expressed in equation (3). However, rather than assuming that there is a stable relationship (θ) between a student's own achievement from year to year, it assumes that there is a stable relationship (ρ) between the average achievement in the same school from year to year regardless of who is attending the school. A useful exercise for future research would be to empirically test hypotheses about these parameters to determine which modeling assumption is more realistic. Another consideration would be whether the variation in achievement between cohorts at the same grade level would be greater than that within a cohort across consecutive grade levels.

Precision of the Cohort Difference Indicator

Even if the bias in the successive cohort indicator could be reduced or eliminated through random assignment of students to schools or careful matching of schools based on student and other characteristics, imprecision could be a major problem. First, assume that true school performance (δ) and student characteristics (X) are nonstochastic. The precision of the cohort difference estimator depends on the variance of the average prior achievement and the average of the unobserved determinants of achievement (e from equation [1]) for each of the two cohorts in the grade of interest, g , as follows.

$$(9) \text{Var}(\hat{\alpha}_{AB}) = \text{Var}(\theta\Delta\bar{Y}_{g-1} + \Delta\bar{e}_g)$$

The variance of the estimator, shown in equation (9), is likely to be quite large because it includes the variance from two cohorts' achievement measures plus "noise" from prior cohorts and prior years that is not related to the school performance in the current period. Assuming that the two cohorts' error terms (e) are independent of each other, the variance of the successive cohort gain indicator can be expanded and written as follows:

$$(10) \text{Var}(\hat{\alpha}_{AB}) = \frac{2\sigma_e^2}{J} \sum_{g=0}^5 \theta^{2g}$$

This is true even in analyses that rely on random assignment. Random assignment statistically equates the average of the treatment and control schools, but all the variation described in equation (10) is still present. The cohort difference indicator amounts to a difference between two cohorts' means, each of which carries along a great deal of unwanted historical information.

Empirical Example

Using data from a large urban district, we constructed three measures of school performance for 51 elementary schools by grade, year, and subject: value-added, average gain, and cohort difference. We used the value-added measure as a standard by which to compare both the cohort difference and the average gain measure. The cohort differences measure always performed worse than the average gain measure relative to value-added.

Our data allow us to look at 2 subjects (math and reading) for 4 grades (two through five) corresponding to 2 school years (three time points: spring 2006, 2007, and 2008), for a total of 16 points of comparison for each estimator. We estimated value-added using equation (1), with the following covariates: gender, income (proxied by free lunch eligibility), race/ethnicity, indicators for over age for grade, disability, and limited English proficiency. We studied the measures themselves in natural units (scale score points) in order to decompose them into their constituent components and we also constructed school rankings based on the three measures because typically school performance measures are used to rank schools.

Correlations with Value-Added

Table 4 shows the correlations of the rankings with the value-added rankings by subject, grade, year, and subject-year. The correlations between the rankings based on the value-added measure and the rankings based on the average gains were greater than 0.90 in all but two of the 16 cases. In comparison, the correlations between the rankings based on the successive cohorts and the rankings based on the value-added measure were less than 0.60 in all cases and less than 0.40 in 10 cases. Overall the correlations of the rankings based on average gains with the value-added rankings were 0.95 in reading and 0.91 in math. In comparison, the overall correlations for successive cohorts were 0.32 in reading and 0.36 in math.⁵

Besides the correlations, we also counted the number of schools that moved 5 or fewer spaces up or down in the rankings relative to the value-added rankings.⁶ This can be seen in figures 1 and 2 for reading and math respectively, by looking at the schools that fall inside of the diagonal lines. In reading across all grades and years, 81 percent schools moved 5 or fewer ranks away from their value-added ranking using averages gains compared to 27 percent of the year-grade-school rankings based on the successive cohorts measure. In math, 75 percent schools moved 5 or fewer ranks away from their value-added ranking using averages gains compared to 28 percent of schools using successive cohorts.

Another way to compare the rankings is to look only at changes in the rankings at the top and bottom of the distribution, given that policy makers may be especially interested in identifying the most and least successful performers in a grade, year, and subject. For this exercise, we ask what would happen if policy makers were using either average gains or cohort differences to identify the top 10 and bottom 10 schools in each grade-year-subject combination, instead of value-added. None of the top 10 reading performers based on the average gain measure are below the median according to the value-added measure. In math, only one percent of all the top 10 average gain performers in each grade and year are below the median according to value-added. For cohort differences, 24 percent of the top 10 performers in reading and 33 percent of the top 10 in math are below the median according to the value-added measure. So the cohort difference measure is more likely than the average gain measure to identify a below-the-median-value-added school as a top ten school. Similarly, the cohort difference measure is more likely than the average gain measure to identify an above-the-median-value-added school as being in the bottom ten.

Correlation of Year-to-Year Changes in Ranks

We also compared whether the measures agreed on whether a school is improving over time. So along with seeing how close the ranks are in a particular year, we looked at whether a school's ranking based on the successive cohorts measure moves in the same direction as its value-added ranking. We counted a change over time as moving in the wrong direction only if the difference was greater than 4 ranks so that we were only counting differences that were more likely to be meaningful. So if the rank according to the average gain measure decreased by two ranks and the value-added rank improved by 2 ranks, this difference does not count as being in the wrong direction. Between 2007 and 2008, the successive cohort reading ranking moves in the wrong direction relative to value-added (by more than 4 ranks) 35 percent of the time, while average gain reading rankings move in the wrong direction only 3 percent of the time. In math, the successive cohort rankings moved in the wrong direction 32 percent of the time, while average gain reading rankings moved in the wrong direction 10 percent of the time.

Decomposition of Bias

Besides using this empirical example to demonstrate the degree to which the rankings based on successive cohorts fail to approximate the value-added ranking, we also want to use it to illuminate the bias decomposition formulas and consider how the components of the bias contribute to the overall biases.

On average, the absolute successive cohorts bias relative to the value-added measure is more than six times as large as the absolute average gain bias (Table 5). The bias associated with successive cohorts measure is measured with less precision than the bias associated with the average gain measure, but the average absolute bias is always highly significantly different than zero for both measures. However the average bias (without taking absolute values) is often not significantly different from zero for both measures. Out of the 16 grade-year-subject combinations, the average gain measure is not significantly different from the value-added measure 13 times at the 95 percent confidence level. The successive cohorts measure is not significantly different 10 out of 16 times.

In terms of the decomposition, we will first discuss the two components that contribute to the average gain bias and then the three components of successive cohorts bias. As discussed earlier, the average gain bias (relative to value-added) has two components. One component is due to the fact that the average gain measure is implicitly assuming that the coefficient on the baseline test score is equal to one. The other component is due to not controlling for background characteristics or implicitly assuming that they have no association with achievement growth (i.e. all the betas equal zero). Either of these components could be negative or positive. So there is the chance for them to cancel each other out. Overall we found a negative correlation between the two components (-0.53 in math and reading). However, the relationship between the two components varies considerably by grade and subject.

The successive cohorts decomposition is more complicated, because there are three components and again either one of them can be negative or positive. The three components in the first level of the decomposition are: (1) the differences in the pretest scores, (2) the differences in the background characteristics and (3) the previous year's value-added score. Whether we are looking by grade and subject or overall, the correlations between these three components are somewhat smaller (between -0.14 and 0.13 in math and -0.40 and 0.32 in reading). The correlations between the sum of any two components with the third component is between -0.13 and -0.16 in math and between -0.18 and -0.38 in reading, so they are less likely to cancel each other out.

For the schools whose rankings suffer from the most severe mis-rankings (either appearing to be very good but actually mediocre or worse or appearing to be very bad but actually decent or better, as described above), the component that contributed the least to the bias tended to be differences due to background characteristics.⁷ This will vary depending on how fast the composition of the student body is changing in a given district and the size of the estimated betas (that is, how much background characteristics influence test scores).

Implications for Policy

Measuring the performance of schools (or districts, states, or programs) is important for accountability and for identifying effective educational practices. This paper has shown that some indicators that may be appropriate for conveying trends in *student* performance can

be flawed indicators of *school* performance. In particular, cohort difference indicators are often used inappropriately for school accountability. When policy makers interpret these indicators as evidence that a school is making effective use of its resources or as evidence that a particular intervention is better than an alternative, they risk making poor resource allocation decisions, replicating ineffective programs, and failing to recognize good teachers and programs that work. For example, ignoring or underestimating the effects of student and family background would cause policy makers to fail to recognize the performance of those teachers who face the greatest challenges and are skilled at working with disadvantaged students. If the flawed indicators are used in an accountability system, they may also discourage teachers from serving disadvantaged populations.

Why Do Good Policy Makers Use Bad Indicators?

It appears that the average gain indicator is potentially unfair, but can be adjusted, while the successive cohort indicator has potential to be quite misleading. If so, then why are they in such widespread use? The reason these indicators are so entrenched in the practice of education policy is a matter for speculation but probably depends on cost and burden as well as on the fact that many student assessments were not designed for evaluation of education interventions. Most were designed instead for diagnosing problems and documenting achievements of individual students without regard to how such achievements were produced.

Another reason that inappropriate indicators are used in evaluations of education interventions may be that researchers unfamiliar with education and student achievement growth are trying to import ideas from the evaluations of welfare, job training, and other programs. It would be important for such evaluation experts to consider the characteristic features of education embodied in the achievement growth model presented here:

- Student achievement is cumulative.
- Family background affects both the level and growth of achievement.
- School effects vary over time and by grade.

The practical implication is that a fair school accountability system or educational program evaluation—that is, one that holds educators accountable for what they can change—must measure student achievement in at least two consecutive grades. In order to hold all grades accountable, it would require some measure of student learning in all grades. In addition, it must acknowledge the role of factors outside the control of school staff, such as the family background of the students.

Endnotes

¹ See *Los Angeles Times*, January 10, 2010 and *Miami Herald*, May 28, 2010 for two more examples.

² This paper uses the terms “test” and “assessment” interchangeably and remains agnostic on the particular instrument best suited to measure student learning. The question of the validity and reliability of the assessment instrument itself is very important, but outside the scope of the current work.

³ In real world settings, students often enter and exit schools during the school year. The proposed model can easily account for such student mobility by allowing I_{ij} to be continuous rather than binary. Thus, for example, if a student transfers from school 1 to school 2 halfway through the school year, both I_{i1} and I_{i2} would take on a value of 0.5.

⁴The value-added approach can be applied at the teacher, school, district, or state level, or at any combination of levels simultaneously, using a multilevel model, sometimes called a “hierarchical” model (Willms and Raudenbush 1989). Similarly, it can accommodate data collected at more than two time points, if such data were available (Willett 1988). For simplicity, the discussion here focuses on two levels, students and schools, with two time points, but the ideas readily generalize to multilevel models.

⁵We focus on comparisons of grade-level measures but also examined school-level measures. A school-level successive cohort measure is expected to approximate the average gain measure more closely because the yearly school cohorts will contain many of the same members. We ranked schools according to a school-level value-added measure (same as grade-level, except included grade dummies), a school-level average gain, and a school-level successive cohort difference. The correlation between school rankings according to value-added and according to average gain was 0.81 in math and 0.92 in reading. For successive cohorts, the correlation with value-added was 0.50 in math and 0.61 in reading.

⁶A move of 5 rankings is an effect size of roughly 0.5 using any of the measures.

⁷This component was less than the component due to pretest differences 91 percent of the time in the math and 98 percent of the time in reading. It was less than the component due to the previous year’s value-added score 91 percent of the time in math and 94 percent of the time in reading.

References

- Bloom, Howard S. (1999). *Estimating Program Impacts on Student Achievement Using “Short” Interrupted Time Series*. New York, NY: Manpower Demonstration Research Corporation.
- Bryk, Anthony S., and Stephen W. Raudenbush. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.
- Coleman, James S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of Educational Opportunity*. National Center for Educational Statistics (DHEW), Washington, DC.
- Cook, Thomas D., and Donald T. Campbell. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin.
- Education Commission of the States. (2001). *Comprehensive School Reform: Programs & Practices*. Available from <http://www.ecs.org/ecsmain.asp?page=/html/issues.asp?am=1>. Most recently accessed April 2001.
- Glazerman, Steven, Susana Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, and Grover J. Whitehurst (2010). “Evaluating Teachers: The Important Role of Value-Added.” Washington, DC: Brookings Institution.
- Greene, William H. (1993). *Econometric Analysis. Second Edition*. Upper Saddle River, NJ: Prentice Hall, 297.
- Kane, Thomas J., and Douglas O. Staiger. (2001). “Improving School Accountability Measures.” NBER Working Paper No. 8156. Cambridge, MA: National Bureau of Economic Research.
- Kohn, Alfie. (2000). *The Case Against Standardized Testing: Raising the Scores, Ruining the Schools*. Westport, CT: Heinemann.
- Ladd, Helen F. (Ed.). (1996). *Holding Schools Accountable: Performance Based Reform in Education*. Washington, DC: Brookings Institution.
- Ladd, Helen F., and Randall P. Walsh. (2002). “Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right” *Economics of Education Review* 21: 1-17.
- McCaffrey, Daniel F., Daniel Koretz, Thomas A. Louis, and Laura Hamilton. (2004). “Models for Value-Added Modeling of Teacher Effects.” *Journal of Educational and Behavioral Statistics*. March 20, 2004. Vol. 29, no. 1. Pp. 67-101.
- Medina, Noe, and D. Monty Neill. (1990). “Fallout From the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America’s Public Schools.” Cambridge, MA: National Center for Fair and Open Testing.

- Meyer, Robert H. (1994). "Educational Performance Indicators: A Critique." Discussion Paper #1052-94. Madison, WI: University of Wisconsin Institute for Research on Poverty.
- Meyer, Robert H. (1996). "Value Added Indicators of School Performance." In Eric Hanushek and Dale Jorgenson (Eds.), *Improving America's Schools: The Role of Incentives*. Washington, DC: National Academy Press.
- Meyer, Robert H. (1997). "Value Added Indicators of School Performance: A Primer." *Economics of Education Review*, 16(3): 283-301.
- Minnesota Department of Children, Families, and Learning. (1999). *Charter School Accountability Framework*. Available from <http://cfl.state.mn.us/charter/accountability.pdf>. Accessed April 2001.
- National Research Council (2011). *A Plan for Evaluating the District of Columbia's Public Schools: From Impressions to Evidence*. Committee on the Independent Evaluation of DC Public Schools. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Pennsylvania Department of Education. (2000). *School Performance Funding Program (SPF) Questions and Answers, November 2000*. Available from <http://www.pde.psu.edu/spfqa.pdf>. Accessed April 2001.
- Raudenbush, Stephen W., and Anthony S. Bryk. (1989). "Quantitative Models for Estimating Teacher and School Effectiveness" In D. R. Bock (Ed.), *Multilevel Analysis of Educational Data*. New York: Academic Press.
- Silva, Elena. (2008). "The Benwood Plan: A Lesson in Comprehensive Teacher Reform." Washington, DC: Education Sector.
- Stone, John E. (1999). "Value-Added Assessment: An Accountability Revolution" In Marci Kanstoroom and Chester E. Finn, Jr. (Eds.), *Better Teachers, Better Schools* (Washington, DC: Thomas B. Fordham Foundation).
- Willett, John. (1988). "Questions and Answers in the Measurement of Change." In E.Z. Rothkopf, (Ed.), *Review of Research in Education Volume 15*. Washington, DC: American Education Research Association; 325-422.
- Willms, J. Douglas, and Stephen W. Raudenbush. (1989). "A Longitudinal Hierarchical Linear Model for Estimating School Effects and Their Stability." *Journal of Educational Measurement*, 26:3, 209-232.

Table 1. Hypothetical Example**A. Average Gains**

Grade	Raw Data			Average Gains (Same Cohort)	
	Year 1	Year 2	Year 3	Year 2–Year 1	Year 3–Year 2
Grade 6	50	40	35	–	–
Grade 7	55	55	55	+5	+15
Grade 8	65	60	59	+5	+4
School Average	56.7	51.7	49.7	+5	+9.5

B. Same Raw Data, Successive Cohort Differences

Grade	Raw Data			Successive Cohort Difference	
	Year 1	Year 2	Year 3	Year 2–Year 1	Year 3–Year 2
Grade 6	50	40	35	–10	–5
Grade 7	55	55	55	0	0
Grade 8	65	60	59	–5	–1
School Average	56.7	51.7	49.7	–5	–2

Table 2. Uses Of Average Gain and Cohort Difference Indicators by Selected States Before NCLB

State	How Test Scores Were Used	Indicator Type
Arizona	School Report Card System includes test score data from multiple grades and years, but until recently the accompanying text typically has interpreted the school's performance as the change from year to year within each grade level. In 2000, the state began including a type of value-added measure for elementary schools. This report card system is the major form of accountability in a state that is a national leader in charter schools.	Successive Cohort Difference
Florida	A May 1999 news release from the Florida Department of Education declared that "teachers, principals, and students are meeting the challenge of higher academic standards with consistently good performance" based on a comparison of the fourth grade average writing test scores in 1999 and 1998.	Successive Cohort Difference
Massachusetts	Statewide testing done in one elementary grade (third grade, changed to fourth grade starting in 2000). In July 1999, the Commissioner of Education released a press report stating that "a majority of Massachusetts elementary schools saw improvement in their third grade reading test scores." A December 1999 press release announced that a Massachusetts foundation awarded \$10,000 each to five principals with the greatest percentage increase in average scores on the statewide test in the previous year. The test score "increases" were differences between the tenth grade achievement levels in 1999 and 1998 for high schools and between the fourth grade levels for elementary schools.	Successive Cohort Difference
Minnesota	The Charter School Accountability Framework allows charter schools in that state to define their own performance measures, but strongly encourages them to use average gains in achievement tests made by the same students from fall to spring of the given year.	Average Gain
Pennsylvania	The Pennsylvania Department of Education gives monetary awards (from \$7.50 to \$35.50 per pupil) to schools based on their "improvement," which is measured as the difference between the current year's cohort and the average of the two previous cohorts in the same grade level (fourth, eighth, and eleventh).	Successive Cohort Difference

Sources: Minnesota Department of Children, Families, and Learning 2000; "Five Principals Recognized for MCAS Improvement," news release, Massachusetts Department of Education, December 22, 1999; "Florida Writes! Scores Show Continuous Improvement," news release, Florida Department of Education, May 10, 1999; "Summary of Performance Funding for Pennsylvania Schools," Pennsylvania Department of Education, June 2, 1999.

Table 3. Summary of Alternative School Performance Indicators

Indicator	Comparison Sample	Data Requirements			Statistical Properties	
		Number of Grades Tested	Units	Control Variables Needed?	Bias ^a	Precision ^b
$\hat{\alpha}$ Value-Added	Same Students	2	Student	Yes	0	$\left(\frac{\sigma_e^2}{(nJ-k)(1-r_{IX}^2)} \right)^{-1/2}$
$\hat{\alpha}_{AA}$ Average Gain	Same Cohort	2	Grade	No	$(1-\theta)\bar{Y}^{g-1} + \beta'X$	$\left(\frac{\sigma_e^2}{J} \right)^{-1/2}$
$\hat{\alpha}_{AB}$ Cohort Difference	Successive Cohorts	1	Grade	No	$\sum_{g=1}^G [\theta^{G-g}(\beta'\Delta\bar{X}_g)] + \sum_{g=1}^{G-1} [\theta^{G-g}(\Delta\alpha_g)] - \alpha_{\beta,G}$	$\left(\frac{2\sigma_e^2 \sum_{g=0}^G \theta^{2g}}{J} \right)^{-1/2}$

^a Bias formulas assume for simplicity that initial differences (in kindergarten achievement) between cohorts are zero and that pre- and post-test data are available for all students where applicable. They also assume that $\text{corr}(I_{ij}, e_{ij})=0$. If this assumption were violated, then each bias term could contain a new component due to omitted variables.

^b Precision formulas assume that restrictions in each model hold and that there are k regressors in the value-added model. For simplicity, they also assume that the covariance of error terms between successive cohorts is zero.

Table 4. Correlations of School Rankings with Value-Added

Grade	Average gains			
	Reading		Math	
	2007	2008	2007	2008
2	0.99	0.96	0.98	0.92
3	0.88	0.91	0.96	0.54
4	0.99	0.92	0.97	0.97
5	0.99	0.98	0.94	0.97
All grades	0.96	0.94	0.96	0.85

Grade	Successive Cohorts			
	Reading		Math	
	2007	2008	2007	2008
2	0.45	0.12	0.58	0.03
3	0.32	0.26	0.39	0.17
4	0.43	0.33	0.48	0.48
5	0.30	0.34	0.22	0.55
All grades	0.37	0.27	0.42	0.31

Figure 1. Comparison of Rankings Produced by Different Estimators, Reading

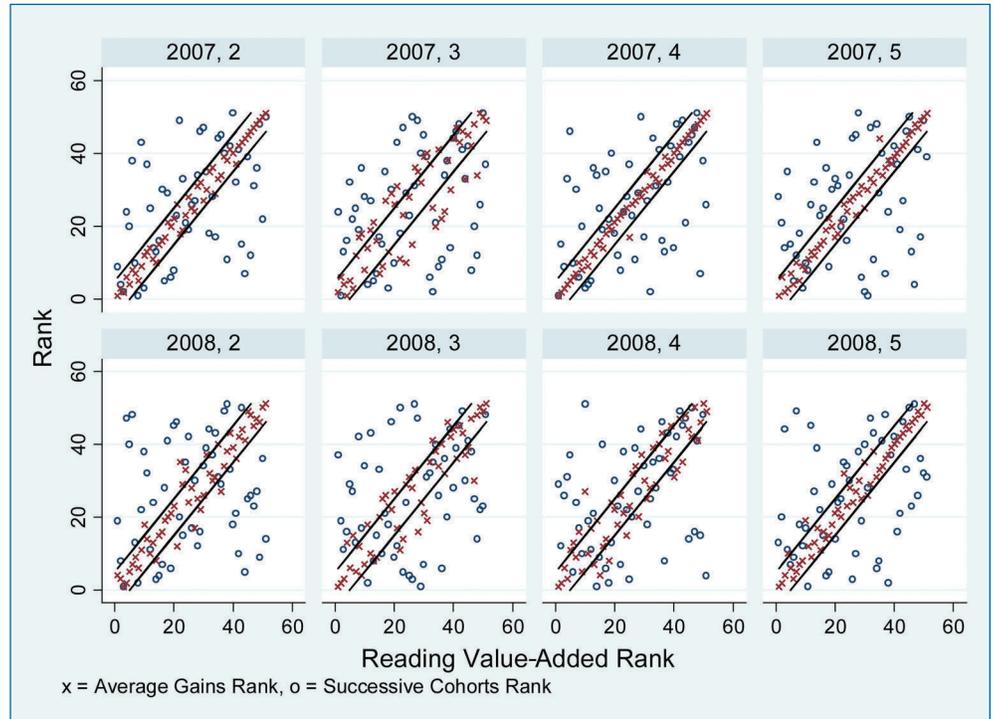


Figure 2. Comparison of Rankings Produced by Different Estimators, Math

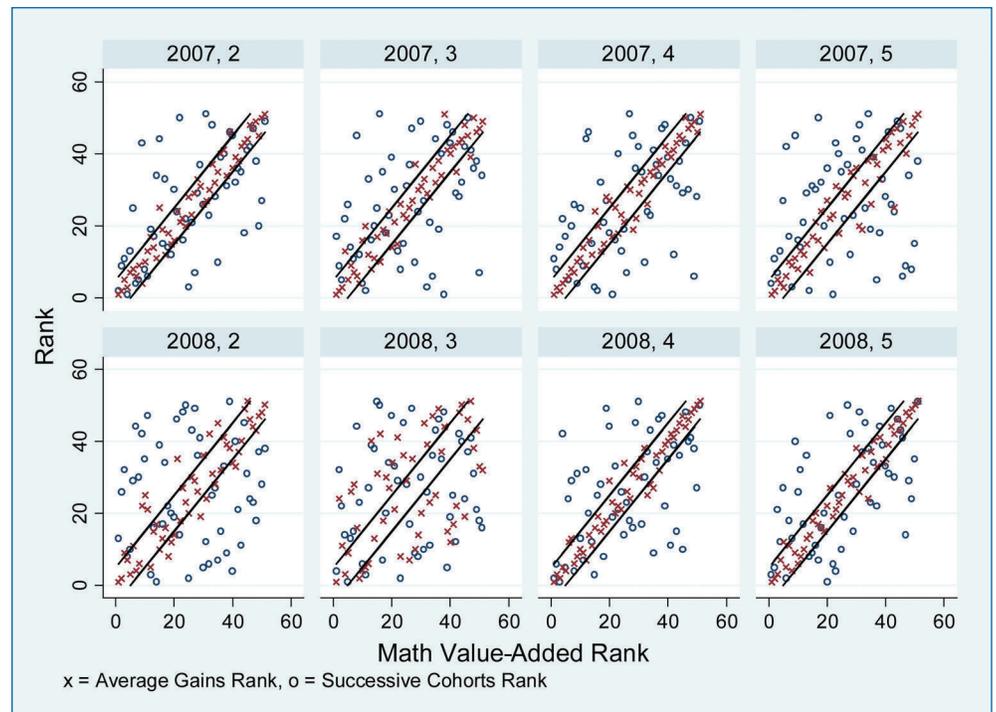


Table 5. Estimated Absolute Bias, by Grade Year and Subject
(Standard errors in parentheses)

Grade	Average gains			
	Reading		Math	
	2007	2008	2007	2008
2	0.01 (0.00)	0.03 (0.00)	0.03 (0.00)	0.04 (0.00)
3	0.05 (0.00)	0.05 (0.00)	0.03 (0.00)	0.12 (0.01)
4	0.01 (0.00)	0.03 (0.00)	0.02 (0.00)	0.02 (0.00)
5	0.01 (0.00)	0.01 (0.00)	0.03 (0.00)	0.03 (0.00)

Grade	Successive Cohorts			
	Reading		Math	
	2007	2008	2007	2008
2	0.12 (0.01)	0.15 (0.01)	0.12 (0.02)	0.18 (0.02)
3	0.15 (0.01)	0.14 (0.02)	0.13 (0.01)	0.18 (0.02)
4	0.13 (0.01)	0.14 (0.01)	0.13 (0.01)	0.14 (0.02)
5	0.15 (0.02)	0.13 (0.01)	0.17 (0.02)	0.16 (0.02)

About the Series

Policymakers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers policymakers and researchers access to our most current work.

For more information about this study, please contact Steven Glazerman at Mathematica Policy Research, 1100 1st Street, NE, 12th Floor, Washington, DC 20002-4221 or by email at sglazerman@mathematica-mpr.com.

www.mathematica-mpr.com

**Improving public well-being by conducting high-quality,
objective research and surveys**

PRINCETON, NJ - ANN ARBOR, MI - CAMBRIDGE, MA - CHICAGO, IL - OAKLAND, CA - WASHINGTON, DC

MATHEMATICA
Policy Research

Mathematica® is a registered trademark
of Mathematica Policy Research, Inc.