Pergamon

Economics of Education Review 21 (2002) 1–17

Economics of Education Review

www.elsevier.com/locate/econedurev

Implementing value-added measures of school effectiveness: getting the incentives right

Helen F. Ladd [a,*], Randall P. Walsh [b]

[a] *Sanford Institute of Public Policy, Duke University, Box 90243, Durham, NC 27708, USA*
[b] *Department of Economics, Duke University, Box 90097, Durham, NC 27708, USA*

Received 1 January 1999; accepted 31 May 2000

Abstract

As part of their efforts to hold schools accountable, several states now calculate and publicize value-added measures of school effectiveness. This paper provides a careful evaluation of the value-added approach to measuring school success with particular attention to its implementation as a tool for increasing student achievement. In practice, even the more sophisticated of the measures currently in use fail to account for differences in resources, broadly defined, across schools and to address the problem of measurement error. The authors find that, as implemented, value-added measures of school effectiveness distort incentives and are likely to discourage good teachers and administrators from working in schools serving concentrations of disadvantaged students. The authors use a large longitudinally-matched data set of fifth grade students in North Carolina to document that approximately two-fifths of the differentially favorable outcome for schools serving advantaged students result from statistical bias associated with measurement error and that correcting for the measurement error leads to significant changes in the relative rankings of schools.

JEL classification: I2

Keywords: Educational economics; Efficiency; Productivity

1. Introduction

As part of the new accountability in K-12 education, states and districts throughout the country are focussing increased attention on student learning.[1] One manifestation of this trend is the proliferation of district- and school-specific report cards, which are designed to provide parents and taxpayers with information on the learning of students. By including information on student outcomes, typically measured by test scores, these report cards use public information as a policy tool to generate pressure for school improvement.

Some states, such as South Carolina, Kentucky, and North Carolina, and some districts, such as Dallas and Charlotte, have gone one step further and offer financial rewards for personnel in schools that appear to be performing well and apply sanctions to schools that are performing poorly. To their supporters, such financial incentive programs are desirable because they induce school officials and staff to focus attention on student learning and serve as a catalyst for change throughout the school system. To their detractors, such programs focus too

* Corresponding author. Tel.: +1-919-613-7352; fax: +1-919-681-8288.

E-mail addresses: hladd@pps.duke.edu (H.F. Ladd), rpwalsh@econ.duke.edu (R.P. Walsh).

[1] See Elmore, Abelmann, and Fuhrman (1996) for a discussion of the new educational accountability which includes a primary emphasis on measured student performance as the basis for school accountability, the creation of relatively complex systems of standards by which data on student performance are compared by schools and by locality, and the creation of systems of rewards and penalties and intervention strategies to introduce incentives for improvement.

0272-7757/01/$ - see front matter
PII: S0272-7757(00)00039-X

much attention on the outcomes that can be measured most easily and, consequently, generate stresses and strains that distort the education process (Bryk & Hermanson, 1992, p. 463; Darling-Hammond, 1992). The overall effect of these programs depends heavily on how they are designed (Ladd, 1996).

Rather than provide a full evaluation of such programs, this paper focuses the measurement of each school's "effectiveness". While clearly not the only design component that deserves attention, accurate measurement of school effectiveness is crucial to the legitimacy and desirability of any school-based accountability system. In order to focus specifically on the design issues associated with measuring a school's effectiveness, we assume that the state (or school district) has successfully developed a political consensus on which parts of its curriculum—most likely reading and math, but possibly other subjects such as social studies and science as well—that it deems most critical and, hence, for which schools should be held accountable.[2] In addition, we assume that for each of the included subject areas the state has a measure of student performance that provides a valid and reliable indicator of students' mastery of the curriculum (see Koretz, 1996 for the difficulties in meeting this assumption).

This research is motivated largely by the recent attempts of states (and districts) to make such measurements in the context of top-down administered accountability systems. However, measures of school effectiveness are also important for other school reform efforts that seek to broaden the choices available to children and their families. For example, in an education system that includes charter schools and parental choice, measures of each school's effectiveness would serve two purposes. One would be to assure that the schools were meeting the public interest that justified their receipt of public funds and the other would be to assure that parents had good information with which to make their decisions among schools.

One of the most promising components of the new educational accountability is the effort by many states to focus on gains in student performance rather than simply the levels of student performance. This value-added approach is promising in that it explicitly recognizes that students who enter a grade with below-average achievement may well leave the grade with below-average achievement, even if school administrators and teachers have made effective use of their resources and the student has made significant gains during the year. Value-added measures have two key elements: (1) they focus on changes in the performance of students from one year to the next (and hence require annual testing of students) and (2) they are calculated for each student within a given school. This latter characteristic allows policy makers to calculate school-specific measures based solely on the students who attended the school for a minimum number of days during the school year. The ability to do so is particularly important for urban schools where the mobility rate of students is typically high.

These value-added measures differ from two other approaches that rely on student test scores to measure school effectiveness. One alternative uses the level of average test scores or pass rates as the measure of school success. Because test scores and socio-economic status are so highly correlated, however, such an approach essentially measures the socioeconomic characteristics of the students in the school rather than the contribution of the school to student learning (see Clotfelter & Ladd, 1996). The other focuses on the rate of each school's improvement during the year (as measured by changes in the test scores of, say, third graders one year to third graders the following year) relative to a school-specific target rate of improvement. Kentucky, which has recently restructured its entire educational governance and finance system, best illustrates this latter approach (Elmore et al., 1996). The main drawback of this approach is that it fails to account for differences in the mix of students from year to year.

As we emphasize in Section 2, a clear distinction needs to be made between school effectiveness measures that focus on overall gains in student performance and those that use an adjusted version of student gains to determine the efficiency with which schools operate. The analysis in this paper focuses on the value-added measures used by South Carolina and North Carolina. Despite the fact that both states treat their measures as if they indicate the efficiency of schools and use them as the basis of incentive awards for school personnel, the measures are best described as indicators of overall student gains rather than of school efficiency.

We have two goals in this paper. One is to contrast such measures to value-added measures that would more accurately measure a school's efficiency. The second is to evaluate such measures on their own terms, that is, as measures of school-specific gains in student performance. We discuss conceptual and practical issues in Section 2. In Section 3, we describe the specific value-added measures used by South and North Carolina and pose a puzzle about why such approaches in practice tend to favor schools serving students from the most advantaged backgrounds. Section 4 partially explains the puzzle based on the analysis of a large matched data set for

[2] At the elementary level, all states with accountability programs include at least math and reading (or language arts). Kentucky has the most ambitious program in that it includes seven subjects, but only at selected grades. In the upper grades, states might want to pay some attention to which subjects are most closely related to labor market outcomes. However, little information is available on this issue (Altonji, 1995).

North Carolina fifth graders. We show in particular that statistical bias arising from measurement error, when uncorrected, explains two-fifths of the correlation between school outcomes and socio-economic status of their children. In Section 5 we highlight unintended incentive effects arising from the use of value-added measures. The paper ends with a brief concluding discussion.

2. Conceptual and practical issues

Economists typically approach the challenge of measuring school effectiveness within the context of a standard education production function. A typical production function might take the following form:

$$A_{it}=\lambda A_{it-1}+\alpha_t \mathbf{S}_t+\beta_t \mathbf{F}_{it}+\varepsilon_{it} \quad (1)$$

where A_{it} equals the achievement of student i in year t, A_{it-1} is that student's achievement in the prior year, $\mathbf{S}_t$ is a vector of school characteristics, $\mathbf{F}_{it}$ is a vector of measurable family background characteristics that affect achievement, and ε_{it} is a random error term (see Hanushek & Taylor, 1990). The lagged achievement term is included to pick up the effects of prior year school and family characteristics. Left out of this standard model are unmeasured characteristics of students, such as their ability and motivation, that affect achievement. Provided such variables have constant effects on achievement over time and that their effects deteriorate at the same rate as prior achievement, they cancel out in this lagged form of the production function.[3]

In order to draw a clear distinction between the resources available to a school and the efficiency with which those resources are used, it is useful to decompose the expression $\alpha_t \mathbf{S}_{it}$ and to rewrite the equation as:

$$A_{it}=\lambda A_{it-1}+\alpha_{Rt} \mathbf{R}_t+E_t+\beta_t F_{it}+\varepsilon_{it}. \quad (2)$$

The vector $\mathbf{R}_t$ is broadly defined to include all factors out of the control of the school's faculty and administrators including budgeted resources, resources provided by parents or foundations, and the composition of the school body, which, through peer effects may affect the learning of others in the class room. E_t measures the effectiveness, or efficiency, of the school's staff and administration. This formulation highlights the fact that one cannot measure the efficiency with which resources are being used without controlling for the resources available to the school.

In practice, however, for a variety of conceptual, practical, and political reasons, it is difficult, if not impossible, for states (or districts) to specify an appropriate vector of family background (**F**) and school resource (**R**) variables. Hence the complete model is never implemented. The first problem is that state agencies typically do not have all of the required demographic data for each student and some of the data they may have is likely to be measured with error. As a measure of family income, the state would probably have to rely on information about whether or not the child is receiving a free or reduced price lunch. While this proxy for income is frequently used by researchers, it is at best an imperfect indicator of family income or other relevant measures of family background. Meyer (1996) suggests an alternative strategy of estimating family characteristics from Census data at the block level. This strategy is appealing in that it could provide a rich set of characteristics including, for example, whether the family is headed by a single person (see Duncombe, Ruggiero, & Yinger, 1996 for evidence of the importance of this characteristic). However, this approach requires that student addresses be known and geocoded. This limitation plus the infrequency of the Census and the residential mobility of many families, especially the families of low-income students, makes this approach impractical.[4]

Second, it is unclear which characteristics of the students should be included. For example, consider the Dallas Independent School District (DISD), whose approach to measuring school effectiveness is in the spirit of Eq. (2). Dallas officials explicitly included as one of the explanatory variables the race of the student (see Clotfelter & Ladd, 1996). The educational logic for including race is not transparent. At best it serves as a proxy for income and family characteristics, such as low income and single parent families, for which other data were not available or were incomplete. In contrast, the political logic for Dallas to control for the student's race in the equation was very clear. Dallas officials wanted to make sure that schools serving minority students had the same probability of being judged an effective school as any other school. The problem is that by applying this criterion of perceived fair treatment, Dallas officials could well have been concealing some true differences in the relative effectiveness of schools serving minorities.

Policy makers in other states have specifically chosen not to control for the race of the student based on political considerations of a different sort. If they were to

[3] See Boardman and Murnane (1979) for other assumptions that would generate this particular form of the production function.

[4] Such an approach is currently used in New Zealand for determining the characteristics of a school's students. Student addresses are geocoded and matched to census mesh blocks which then are the source of information for five socioeconomic characteristics of the students. This approach works better in New Zealand than it would work in the United States since the Census is conducted in that country every 5 years (Fiske & Ladd, 2000).

include race as a control variable, they faced the possibility that they might be misinterpreted as sending a signal that the academic expectations for minority children are lower than those for white children. Such a message would be inconsistent with the rhetoric that underlies much of the outcomes oriented reform efforts, namely that all children can learn to high levels. While this concern about a specific demographic variable applies most pointedly to a student's race, it applies as well to other background characteristics of students, such as family income.

A third problem is that the endogeneity of the school composition variables makes it difficult to control for peer effects. For example, the inclusion in the equation of measures such as the average test performance of all children in the school (or preferably, in the relevant grade in the school) or the percentage of children from economically disadvantaged families, creates a potentially serious problem of endogeneity. This problem stems from the observation that, for at least some students, the peer group is a factor in the family's decision about what school the child will attend, either through its choice of neighborhood or, in a district in which there is some school choice, by its explicit choice of a school. Other researchers have demonstrated the importance of this endogeneity in the context of models of teen pregnancy and the decision to stay in school. Not accounting for the endogeneity, Evans, Oates, and Schwab (1992) find evidence of peer effects with respect to both forms of behavior. However, when they account for the endogeneity, they find no peer effects. We believe these estimation issues are serious. Accurately estimating the peer effects is hard and would require a much richer data set than would typically be available to state agencies.[5]

Finally, controlling for resource levels is also far more difficult than one might expect. Surprising as it may sound, states typically have little or no information on the resources available to individual schools. Only a few states such as Texas and Ohio routinely collect that information. In general, the state maintains data on resources and spending only at the level of the school district.[6] An additional potential challenge arises in sorting out which components of a school's resources are under its control and which are not. For example, to the extent that a school raises additional funds through either its own entrepreneurial activity or because of the high quality of its programs, the additional funding should not be viewed as outside the control of the school.

The bottom line is that the full model as specified in Eq. (2) is not being implemented. The one district, Dallas, that has gone the furthest in implementing it, falls short by not including family background variables (other than those measured by eligibility for subsidized lunch or proxied by race), by not including resource variables (other than through a school crowding variable) and not controlling for school mix effects (Clotfelter & Ladd, 1996).

Some states, such as South and North Carolina, have opted for simpler value-added approaches that require much less data.[7] The spirit of the South Carolina approach is represented by the following equation, where test scores are the measure of student achievement, j signifies the school and E'_{jt} represents a measure of school effectiveness (see Section 3 for the actual approach that does not include school indicator variables):

$$\text{Test score } t_{ijt} = f(\text{Test scores}_{i,j,t-1}) + E'_{jt} + \varepsilon_{ijt}. \quad (3)$$

The spirit of the North Carolina approach is captured by the following specification (see Section 3 for the specific functional form):

$$\text{Test score}_{it} - \text{Test score}_{i,t-1} = f(\text{Test scores}_{i,t-1}) + E'_{jt} + \varepsilon_{it} \quad (4)$$

Note that neither of these approaches includes any variables other than student test scores. Because of the high correlation between student test scores and student background, the inclusion of prior year test scores controls for some, but not all, of the effects of a child's family background. No controls are included for school resources.

Importantly, the exclusion of the school resources changes the interpretation of the school effects compared with the full production function model. The estimated school effects in this equation should be interpreted as all the effects on student learning associated with each school, including both those that are within the control of school personnel and those that are not. Hence, they

[5] Specifically, consistent estimation would require a vector of demographic data for each student's family such as the parents' income and educational background to predict how the student and her parents chooses her peers. Further discussion of the issues associated with estimating this type of social effect are contained in Manski (1993).

[6] This statement applies as well to the federal government. Currently the National Center for Education Statistics collects information on resources only at the district level but is currently under pressure from Congress to develop an approach for collecting it for individual schools.

[7] South Carolina was the first state to introduce a formal school based accountability system. The system that we are describing here was introduced in 1984 and is now being changed. North Carolina introduced its program for the 1996–97 school year. See Clotfelter and Ladd (1996) for a description of the South Carolina program. North Carolina's program is described further below and in North Carolina State Board of Education (1996b).

do not measure how efficiently the school is being operated.[8]

The question then becomes, can this simpler approach be justified? From some perspectives the answer is yes. First, the estimate of school effectiveness that emerges from this model provides useful information for the typical parent and student in their capacity as choosers of schools. Whether parents are choosing schools by choice of neighborhood as part of the residential location decision, or by choice of school within a system that offers public school choice to parents, this measure provides them with information about how one school compares with another in terms of its ability to add to the learning of its students.

The measure also provides information to parents, citizens, and policy makers in their role of trying to assure that all schools contribute to student learning. Poor performance by this measure would indicate that something needs to be done to make the school more effective in increasing the learning of its students. Importantly, however, what may be needed is not just harder or smarter work by existing teachers. Instead, major interventions, such as investment of additional resources, may be needed to counter the effects of other factors, such as the school's mix of students, outside the control of school personnel.

Because it does not control for school resources, broadly defined to include the mix of a school's students, that are outside the control of the schools, the measure of school effectiveness that emerges from this approach should at best be used with caution as the basis for rewards and sanctions for principals and teachers. Only if all schools had adequate resources that fully accounted for the mix of students they serve would it be fair and appropriate to use this measure of school effectiveness for that purpose. If this requirement is not met then the teachers and principals in schools serving students from disadvantaged backgrounds would be inappropriately penalized for factors outside their control. As a result, such schools would find it difficult to recruit high quality teachers. The accountability program would create incentives for these teachers to shun such schools in favor of other schools where they had a greater chance of being rewarded and a smaller chance of being sanctioned.

Nonetheless, this measure could still drive policy in productive directions. That would occur, for example, to the extent that it put pressure on policy makers to try to determine the causes of school ineffectiveness and to intervene in ways to make the low-performing schools more effective. If a state plans to use measurement approaches of this type, it behooves the state to develop the best possible measures with the available data. In the following two sections, we describe the models used in South and North Carolina and then examine whether their models pass this test.

[8] This point has also been made elsewhere. See, for example, Meyer (1996).

3. Puzzles from South Carolina and North Carolina

An important stylized fact characterizing the value-added measure as it has been implemented in South Carolina and North Carolina is that schools serving higher performing students are more likely to be deemed effective than schools serving low-performing students. One apparent explanation for this result springs to mind: the students who performed well during the prior year are likely to learn more during the year than students who performed poorly. Indeed, their high prior-year scores probably reflect above-average annual gains in the past. However, this does not appear to be the explanation in either state. It does not explain the South Carolina result because that state models the relationship between current and prior year test scores as a nonlinear relationship. If low-performing students typically learn less in a year than high-performing students, that fact will be incorporated into the prediction equation and will not show up in the estimate of a school's effectiveness. North Carolina tries to account for that possibility in a different way, namely by including a factor in the equation to account for the faster learning of more proficient students.

Thus, there is a puzzle. Three other explanations are worth considering. The first is that the schools serving high-performing students (who typically are the more advantaged students) may indeed be more effective than the schools serving low-performing students. This explanation is plausible for several reasons: compared with other schools, such schools may have more resources, they may generate greater positive peer effects, and they may attract higher quality teachers who can use their seniority to move to schools where the students are more motivated and easier to teach. The other two explanations are statistical: the equations could be mis-specified or they could be subject to measurement error. In Section 4 we use data from North Carolina to examine these two statistical explanations and show that approximately two-fifths of the correlation between student background and school effectiveness arise from failure to correct for measurement error in the test data.

Before moving to that section, we spell out in more detail the South Carolina and North Carolina approaches. The South Carolina approach is the more straightforward. Using a combination of tests (a nationally normed test in some grades and a state criterion-referenced test in other grades), South Carolina predicts a test score for each student in each grade for each of the basic subjects, such as math, using the following equation:

$$M_{it}=\alpha+\beta_1 M_{i,t-1}+\beta_2 R_{i,t-1}+\beta_3 M^2_{i,t-1}+\beta_4 R^2_{i,t-1} \quad (5)$$

$$+\beta_5(M_{i,t-1}R_{i,t-1})+\varepsilon_{it}$$

where M stands for the student's test score in math and R her score in reading. The parameters α and β_1–β_5 are estimated by regression analysis on the basis of all students for whom both current and prior-year test data are available.[9] Rather than estimate school effectiveness for each grade and subject by including school-specific indicator variables, South Carolina uses Eq. (5) to predict a test score for each student and then defines the student's gain as the difference between the student's actual score and her predicted score, that is, the residual from the estimating equation. The school's gain index (SGI) is then calculated as the median of the student gains across all students in the school.[10]

Two things are worth highlighting about this approach. First, the schools are ranked relative to each other rather than relative to some school specific target rate of growth. Consequently, if one school moves up in the ranking, another school must move down. Also, a school's ranking in any one year depends not only on how effective the school was during that year by this measure but also on how effective the other schools were during that year. Second, effectiveness is measured relative only to other schools within the state; an effective school by South Carolina standards may not be viewed as effective when compared with schools in other states.

As has been documented elsewhere (Clotfelter & Ladd, 1996), this approach tends to rank schools serving students with high SES more highly than those serving students with low SES. Because South Carolina officials believed it would be politically unacceptable for any ranking system to have a clear bias of that type, they divided all schools in the state into five clusters, defined primarily by the socioeconomic characteristics of the students they served. In this way, schools competed only with schools within their division for the awards that were given to the top 25 percent of the schools in each division.[11]

Because it was developed more recently, the approach used by North Carolina more fully reflects the ideas of the standards movement and systemic reform. Systemic reform calls for curriculum frameworks, that is, a clear statement of what the state wants children to know and be able to do, assessments that test a student's mastery of that framework, and relatively clear targets for schools to aim for. Since the 1992–93 school year, North Carolina has been administering end-of-grade tests to assess student mastery of the state's curriculum in reading and math. Because the scores from these tests are reported on developmental scales, they provide a ruler for measuring growth across time, and hence across grades (North Carolina State Board of Education, 1996a). By looking at past average growth in, for example, fifth grade math, the state can define a year's worth of learning. Schools then are expected to generate "a year's worth of learning for a year's worth of work". Schools that do so are recognized as being effective (or exemplary if they exceeded their growth target by more than 10 percent) and others are either not recognized or put into the category of low-performing schools. To be labeled low performing, a school must not only fail to meet its growth target but must also have less than half of its students performing at or above grade level.[12]

Our interest here is how the state determines the growth standards which vary by grade and by subject. Consider, for example, fifth grade reading. Using data for all the students for whom the state could match reading scores from fourth to fifth grade in 1994, the state estimated an equation of the form:

$$\text{Test score}_{it}-\text{test score}_{i,t-1}=\alpha+\beta_1 X1_{i,t-1}+\beta_2 X2_{i,t-1} \quad (6)$$
$$+\varepsilon_{it}$$

where test score$_{it}$ refers to a student's test score in fifth grade reading in the current year (in this case, 1994), $X1$ is intended to measure that student's proficiency and $X2$ to account for the possibility of regression to the mean (and will be defined more precisely below). The idea

[9] Throughout the rest of this paper we use this quadratic version of the estimating equation. However, we return to the issue of functional form below.

[10] South Carolina policy makers chose to use the median rather than the mean so that teachers would not have an incentive to focus attention on the students likely to gain the most while ignoring the others. However, the use of the median allows teachers to ignore both ends of the distribution. In implementing its conceptually similar school accountability system, in contrast, Dallas chose to use the mean of the student scores so that teachers would have to pay attention to the learning gains of all students. Thus, whether the mean or the median is preferred cannot be determined on technical grounds alone. The preferred alternative depends both on behavioral responses that are not yet fully understood and on policy makers' values.

[11] In response to a number of complaints about this clustering approach, the state introduced in 1992 an alternative ranking system, referred to as exceeding expectations, which compared the size of a school's actual gain with that predicted based on the absolute level of scores in that school. A school could win a reward either by being in the top quartile of the socio-economic group into which it was placed or by being in the top 25 percent of the distribution of schools in the degree to which it exceeded expectations (Clotfelter & Ladd, 1996, p. 32).

[12] The performance standard is calculated by taking a weighted average across all grades for each performance level (1=below grade level, 2=at grade level, 3=grade level, 4=highly proficient).

here is that the change in test score for each student is expected to depend on the average change in test scores for all students (as measured by α) adjusted for the student's proficiency level and the possibility that the student's 1993 fourth grade score may have been relatively high or low because of a large random error that would disappear the following year.

Adjusting for proficiency accounts for the realistic possibility that more proficient students are likely to learn more in a year than less proficient students. In the absence of a measure of a student's true proficiency, the state approximates it by the sum of the student's fourth grade reading and math scores. To simplify the interpretation of the coefficient as the effect of a deviation from average proficiency, $X1$ is defined as the sum of the reading and math score minus the state average reading and math scores.

The variable $X2$, which is designed to account for the possibility of regression to the mean, is the student's fourth grade reading score expressed as a difference from the state average score. A negative sign is predicted for the coefficient of this variable. A student with an above-average score in the fourth grade may experience below average growth in the fifth grade simply because some of the fourth grade score may represent a large positive random error. Analogously, a student scoring below average in the fourth grade may experience greater than average growth in the fifth grade.

Based on a matched group of approximately 50 000 students, the estimated values for b_0, b_1, and b_2 for fifth grade reading in 1994 were 4.5, 0.21, and -0.60 and for fifth grade math were 7.3, 0.22 and -0.57 (Sanford & Thissen, 1995, p. 7). These estimates indicate that the average gain for the typical student on the development scale in fifth grade reading was 4.5 points and for math 7.3 points. For a student with above-average fourth grade scores, there would be two partially offsetting adjustments: the student's higher proficiency would lead to a higher expected gain and regression to the mean would reduce the expected gain. The net effect, based on the estimated coefficients, would be to reduce the expected growth rate for this student by approximately 0.18 times the difference between the student's fourth grade score reading score and the average and by 0.13 times that difference in math.[13]

So that schools will have relatively clear and stationary targets to aim for, the targets will continue to be based on the equations from the 1993–94 year until the State Board of Education decides to change the base year.[14] Based on the experience in that year, the average rates of growth (α), and hence the expected gains, for each grade in each subject are higher in the lower grades than in the higher grades. Although one would expect the estimated values of the adjustment parameters (β_1 and β_2) to vary from grade to grade and subject to subject, the observation that the estimated values of β_1 and β_2 did not vary much across grades encouraged the state to apply adjustment coefficients that were uniform across grades.[15]

Because it is linear, the equation that predicts expected growth applies to the averages of test scores as well as to the scores of individual students. Hence, given the prior year test scores of their students, each school, with the help of a computer package provided by the state, is able to calculate its own targets at the beginning of the year and at the end of the year to determine how the school has done relative to its targets.

As noted at the beginning of this section, this approach generates the result that schools that serve higher SES or higher performing students are more likely to meet their targets than schools serving lower SES or lower performing students. For the first year of the program, the average percentage of students not eligible for free and reduced price lunch (a measure of SES) was higher in schools that met or exceeded expected growth (67.5%) than in schools that did not meet expected growth (59.8%). The average percentage of students reading at or above grade level was 73.1% in schools that met or exceeded expected growth and was only 61.9% in schools that did not meet expected growth. Similarly, large differences in levels were found for proficiency in math (77.5% vs. 63.4%).

4. Statistical considerations in measuring value added

In this section, we use a rich 3-year data set on student test scores in North Carolina to explore how two statistical problems—specification error that arises because their value-added equations do not include school-specific indicator variables and measurement error—affect

[13] This estimate is based on the simple (but not unrealistic) assumption that for a particular student math and reading scores tend to be about the same. Since $X1$ is the sum of the reading and math score, the net effect is approximately $2\beta_1+\beta_2$.

[14] Note, the parameters of the North Carolina model can be recovered from a simple OLS regression of, for instance, fifth grade math scores on fourth grade math and reading scores. The crucial distinction between the North Carolina approach and a simple OLS model is the assumption that the coefficients in Eq. (6) are constant across time. Thus, under the NC interpretation, Eq. (6) can be updated each year without any additional statistical work. Instead, new state-wide test averages are combined with previously estimated values for the parameters α, β_1, and β_2 to arrive at a given year's formula for predicted test scores.

[15] For reading, $\beta_1=0.22$ and $\beta_2=-0.60$. For math, $\beta_1=0.26$ and $\beta_2=-0.58$.

the value-added measures used by South Carolina and North Carolina. Empirically, measurement error turns out to be the more important problem by far. As we show below, correcting for measurement error substantially changes how the schools are ranked and, in particular, raises the effectiveness measure for many schools with low average test scores and reduces it for many schools with high average test scores. Hence, failure to adjust for measurement error significantly biases the measures of school effectiveness against the schools serving low-performing students.

4.1. The data

We focus for simplicity on measuring a school's value added in the fifth grade only. While a more complete measure of school effectiveness would include the value added in all of the grades offered by the school, the patterns based on fifth grade scores suffice for illustrative purposes. To measure value added in the fifth grade, we need fifth grade scores in one year and fourth grade scores for the same students in the previous year. In addition, third grade scores for the same set of students are needed to either to correct for measurement error or, depending on the assumptions of the model, to specify the full model.

Our data set includes test scores in math and in reading matched over 3 years for more than 37 000 North Carolina students. The data cover all of the students in the state for whom we were able to match third, fourth, and fifth grade scores over the period 1993–95. Because we did not have access to student identifiers, we matched students by birthday, gender, and school.[16] For example, consider a third grade female student in 1993 in a specific school. That student's test scores remained in the sample provided we were able to uniquely match that student to a girl with the same birthday in that same school in each of the following 2 years, that is, when she was in fourth and fifth grades. Our matched sample represents about 44 percent of the statewide population of students taking the end-of-grade tests.

This matching process is imperfect in that we lose all students who moved from one school to another during the 3-year period and all students with non-unique criteria for matching (e.g. twin boys in the same school). The latter limitation is not very serious and should not bias the sample. The loss of the movers, however, leads to a sample with higher average test scores and a lower proportion of minorities than the entire population. Appendix Tables A.1 and A.2 indicate the magnitude of the sampling effect by comparing our matched sample with the population of all test-taking students. The bias toward white students and students with higher test scores in the matched sample should have little or no effect on the analysis presented below.[17]

4.2. Specification error related to the indicator variable

If schools do indeed differ with respect to their effectiveness in increasing the learning of students, then a value-added model for fifth grade learning should include school-specific indicator variables to pick up the school effect. The standard approach would be the fixed effects models illustrated by Eqs. (5) and (6). However, neither North nor South Carolina included the school specific indicator variables in their model. Instead, they measure school effectiveness as the median of the differences between actual test scores and predicted test scores (in South Carolina) or as the mean of the residual differences between actual and predicted changes in test scores.[18]

We begin with the South Carolina model and, for simplicity, we measure each school's effectiveness by its contribution to fifth grade scores alone. The fifth grade scores are predicted as a function of fourth grade scores, with the estimated residual for each student—that is, the difference between the student's actual and his or her predicted score—measuring each student's gain. In practice, South Carolina uses the median of the student gains to aggregate them to the school level. We simplify by

[16] Ideally, the student's race could also be used in the matching process. Unfortunately, North Carolina changed the category definitions used for recording race in the middle of the sample period, rendering unusable the racial identifier for the matching process. Based on comparisons between matches for 2 years for which we had consistent measures of a student's race, we found that the overall impact of excluding race from the matching process was small and inconsequential.

[17] Student identifiers would have helped but would not have eliminated completely the bias in the sample. Sanford and Thissen (1995) had access to student identifiers since they were under contract to the state. We compare their 2-year matched data to our results using only 2 years of matched data. The average fourth grade reading test score in their sample is 148.0 with a change of 5.0 from third grade to fourth grade, while ours is 148.66 with a change of 5.05. For math, their average fourth grade test score is 147.3 with a change of 7.0, while ours is 148.17 with a change of 7.17. Also note that in measuring the effectiveness of schools using either matching method, any students leaving a school in the middle of the year or coming to a school from out of state at any time would be excluded from the measure of a school's effectiveness. This fact should actually reduce any bias resulting from the data matching process.

[18] As noted earlier, the predicted test scores in North Carolina emerge from equations based on data for a prior period so that schools know at the beginning of the year approximately what their goals are.

using the mean rather than the median as the measure of a school's effectiveness.

The question is how much error emerges in the ranking of schools when mean residuals rather than fixed effects are used to measure school effectiveness. In particular, assume that the true model of student performance on fifth grade math tests is given by an equation similar to Eq. (5), modified to allow the intercept term to vary from one school to another. In that case, the estimated value of the intercept for each school would provide an estimate of the school's effectiveness. Note that if the equation is estimated without the school-specific intercepts and school effectiveness is measured as the mean residual, the estimated measure of school effectiveness will in general be biased. For example, to the extent that higher-ability students tend to cluster in schools that are more effective (that is, they would have higher intercepts in the fixed effects model) and lower-ability students cluster in schools with lower intercepts, the residuals approach will rank schools whose average student ability is at the low end too high and those whose average student ability is at the high end too low relative to their true ranking. However, which way the bias goes is an empirical question that depends on the correlations in the sample.

The top panel of Table 1 illustrates that for our sample the bias is consistent with the example just given, but that the bias is small. The table requires some explanation. The entries report how school effectiveness rankings change as we vary the model from one form to another. We begin by calculating a ranking for each school for each model by assigning each school an integer from 1 to 10 based on the decile in which the school's effectiveness measure falls—1 is the least and 10 is the most effective. Our focus on relative rankings reflects South Carolina's emphasis on the effectiveness of each school relative to other schools. We report the changes in rankings as the percentage of schools in a given group whose ranking moves from one decile up to a higher decile or down to a lower decile.[19] This is done for a set of four reference groups, based on third grade test scores in math.[20] Group 1 includes the 10 percent of schools serving students with the lowest average third grade performance, group 2 includes the bottom half of the distribution, group 3 includes the top half, and group 4 the highest 10 percent of the schools.

The first panel shows that the shift from the mean residuals model (SC SGI) to the fixed effects model would decrease the rankings of 5.2 percent of the schools in group 1 and of 3.4 percent of the schools in group 2. Analogously, the shift to the fixed effects model would raise the rankings in 3.9 percent of the schools in group 3 and 9.1 percent of the schools in group 4. Very similar patterns arise for the shift to the fixed effects specification of the North Carolina model in the bottom panel.[21] Of interest is that not only are the shifts small but that they serve to exacerbate, and not to explain, the puzzles for the South and North Carolina data that we outlined earlier. Hence, we must look elsewhere for a possible explanation for the puzzle.

4.3. Measurement error

In principle, tests are designed to measure how much students know. In fact, however, they measure how much students are able to show what they know in the form of test answers. For example, two students may know the same amount, but if one is a better test taker than the other, then the better test taker will have a higher true test taking ability even though she has no more knowledge of the subject matter. Clearly, actual test scores are imperfect measures of a student's true knowledge. However, they are also imperfect measures of true test taking ability. For a variety of reasons, a student could have a bad day or a good day at any particular sitting of the test. That means that the student's actual test score is the sum of her true test taking ability and a random error. Thus, test-taking ability is measured with error.

When the error is in the dependent variable (in our case, fifth grade test scores), it creates no statistical problem; the random component simply shows up in the error term of the regression. More problematic is when the explanatory variables—in our case, fourth grade scores—are measured with error. The problem arises because the regressor is correlated with the error term.[22] This violates one of the assumptions necessary to obtain unbiased estimates using ordinary least squares. As is well known in the context of a model with a single regressor, the estimated coefficient will be biased toward zero. With multiple regressors, all the estimates will be biased but in unknown directions.

A common approach to eliminating measurement error

[19] The focus on movement from one decile to another decile is designed to capture significant movements in a school's relative ranking as opposed to small and less significant changes such as that which would occur when two schools with adjacent relative rankings switch places.

[20] We use students' third grade scores rather than their fourth grade scores as a measure of student ability to avoid any potential correlations in the error with which we measure ability and in the changes reported in the table. Having said this, sensitivity tests showed that the results presented in Table 1 are robust to the use of either third, fourth or fifth grade scores and either reading or math scores.

[21] A model which allowed for school specific slope coefficients was also tested, yielding similar results.

[22] See Greene (1993) for a detailed description of measurement error.

Table 1
Effect of adopting an 'improved' model specification on the decile ranking of schools in different reference groups[a]

	Group 1 (low 10% on 3rd grade math)		Group 2 (low 50% on 3rd grade math)		Group 3 (high 50% on 3rd grade math)		Group 4 (high 10% on 3rd grade math)	
	Decile improved (%)	Decile dropped (%)	Decile improved (%)	Decile dropped (%)	Decile improved (%)	Decile dropped (%)	Decile improved (%)	Decile dropped (%)
South Carolina								
SC SGI[b] to SC Fixed Effects[c]	0.0	5.2	0.0	3.4	3.9	0.5	9.1	0.0
SC SGI[b] to SC Instrumental Var.[d]	61.0	7.8	38.2	10.7	12.3	34.7	2.6	59.7
SC SGI[b] to SC F.E. and I.V.[e]	62.3	9.1	38.5	11.0	12.3	34.7	2.6	59.7
North Carolina								
NC OLS[f] to NC Fixed Effects[g]	0.0	7.8	0.8	4.7	4.2	0.3	7.8	0.0
NC OLS[f] to NC Instrumental Var.[h]	51.9	6.5	37.2	9.7	9.7	36.8	0.0	67.5
NC OLS[f] to NC F.E. and I.V.[e]	49.3	7.8	36.6	10.2	10.5	36.4	1.3	66.2

[a] Owing to computational constraints associated with estimation of the combined Fixed Effect and Instrumental Variable models, the results reported in Table 1 are based on all matched test data for a randomly chosen sub-sample of 763 elementary schools. With the exception of the two F.E. and I.V. models, all calculations were also done using the entire sample of 997 schools. There are no substantial differences between the two samples.

[b] The SC SGI (Student Gain Index) model measures school effectiveness as the mean residual from the regression of fifth grade math scores on fourth grade math, fourth grade reading, fourth math×fourth reading, fourth math squared, fourth reading squared, and an intercept.

[c] The SC Fixed Effects model is identical to the SC SGI model except that school specific intercepts (fixed effects) are included in the model and replace the mean residuals as the measure of each school's effectiveness.

[d] The SC IV model is the same as the SC SGI model except that third grade reading and math scores are used as instruments for the fourth grade scores.

[e] The S.C. and N.C. F.E. and I.V. models include both model improvements in a single model estimation.

[f] The NC OLS model measures the average residuals from a regression of (fifth grade math (reading)−fourth grade math (reading)) on the difference from the mean of (fourth grade math+fourth grade reading), the difference from the mean of fourth grade math (reading), and an intercept.

[g] The NC Fixed Effects model is identical to the NC OLS model except that school specific intercepts (fixed effects) are included in the model and replace the mean residuals as the measure of each school's effectiveness.

[h] NC IV uses third grade math and reading scores as instruments in the above model.

is the method of instrumental variables. This approach removes the correlation of the regressor with the disturbance term by using an instrument that is correlated with the regressor, but not correlated with the error term, yielding consistent parameter estimates. Finding such a variable is the key to this approach. We solve the potential measurement problem by using third grade test scores as instruments for fourth grade scores.

The use of third grade test scores as instruments requires that the third and fourth grade test scores be correlated. This requirement is clearly met. The auxiliary equations, which predict fourth grade math and reading scores as functions of third grade math and reading scores, generate an R^2 of 0.70 and 0.71 for math and reading, respectively. Additionally, because third grade scores are also measured with error, we require that measurement error in the third grade test scores be uncorrelated with the measurement error in the fourth grade test scores. As discussed above, these scores measure the ability of students to perform on a test of a given style and scope as chosen by the state education agency. The likely sources of measurement error are factors associated with a particular student's performance on a given day such as fluctuations in the student's mood or focus that would not be correlated across test taking dates.[23] Factors which vary systematically across students are

[23] One possible exception to this characterization would be systematic and persistent differences in the test-taking environment experienced by individual students. The authors believe that because of the change in grade, teacher, and classroom between testing dates, and the strong incentives in place for schools to provide optimal testing conditions for their students, that this source of error is small relative to the student-specific fluctuations.

components of the student's test taking ability and would not enter the error term. For these reasons, third grade test scores make natural instruments for solving the measurement error problem.

Using these instruments, we now ask is measurement error a problem? Following Hausman (1978), we use a Wald test to check for the presence of measurement error. The test provides clear and convincing evidence that measurement error is a potentially serious problem. For example, the chi-squared statistic for math using the South Carolina model is 2508, which far exceeds the critical value of 12.59.[24]

As shown in Table 1, correcting for measurement error greatly affects the school value-added rankings as defined above. Consider for example the second row which refers to rankings based on math scores using the South Carolina approach. The entries indicate that using instrumental variables to correct for the measurement error raises the decile rankings for 38 percent of the schools in group 2 (the bottom half of the ability distribution) and lowers the decile ranking for 35 percent of the schools in group 3 (the top half of the distribution). The systematic nature of the bias is even more pronounced at the tails of the distribution. The rankings of about 61 percent of the schools in group 1 (schools serving students with the lowest 10 percent average third grade scores) would have been higher based on the IV model and the rankings of about 60 percent of the schools in group 4 (highest 10% of ability) would have been lower.[25] Thus, failure to correct for measurement error leads to the incorrect classification of many schools and the bias is against those serving low-performing students and in favor of those serving high-performing students. The third row, which summarizes the combined effects of the shift to a fixed effects model and the shift to instrumental variables, presents a similar overall picture.

Thus, measurement error provides some of the explanation for the puzzle that emerged in South Carolina. Had South Carolina policy makers correctly adjusted for measurement error, many of the schools serving low-performing students (which also tend to be those serving students with low socioeconomic status) would have been declared more effective than they appeared to be according to the state's ranking and the reverse would have been true for schools serving high-performing students.

Not surprisingly given its similarity to the South Carolina approach, measurement error also emerges as a statistical problem with the North Carolina approach. The second row in the North Carolina panel of Table 1 shows that the correction for measurement error changes the rankings in ways that are comparable to the changes under the South Carolina approach. Specifically, almost 52 percent of the schools in group 1 would have had a rise in their rankings (by at least one decile) and about 68 percent of the schools in group 4 would have had a fall in their rankings. The bottom row summarizes the combined effects of the two model adjustments.

4.4. Correlations between school effectiveness rankings and grade or school characteristics

Table 2 illustrates the correlation between school effectiveness measures (as estimated separately based on math and reading scores) and average student performance and the economic and racial composition of the students. The first row for each subject simply documents the well known fact that average test scores are negatively correlated with the share of a school's students who are economically disadvantaged or are African-American. Clearly, if the measure of school effectiveness were average test scores, schools serving disproportionately poor or minority students would typically look ineffective regardless of their value added.

Use of either the South Carolina or the North Carolina model for calculating a school's value added would still yield a measure of school effectiveness that is positively correlated with average performance and negatively correlated with the percent of students eligible for subsidized lunches or the percent black. Thus, the North Carolina data confirm the puzzle identified by policy makers in South Carolina. Even when one bases a school effectiveness measure on the differences between actual and predicted scores, schools that have a disproportionately high intake of high income and white students continue to look better than those serving students from more economically or racially disadvantaged backgrounds.

The models that correct for both the specification error and the measurement problem (labeled F.E. and IV in the table) show that about two-fifths of the correlation

[24] Repeated Hausman tests were performed under various specifications. The tests consistently yielded test statistics in excess of 2000.

[25] To rule out the possibility that these systematic changes in the deciles are the result of instability introduced by the instrumental variables procedure, we explored how the rankings would change if a random error with known variance was added to the initial school scores. When the SGI is perturbed using an error term with mean zero and variance equal to the variance of the SGI, 43 percent of the group 1 schools see an improvement in their decile and 38 percent a decline. The changes are: 40 percent up 36 percent down, 40 percent up 40 percent down, and 43 percent up 34 percent down for groups 2, 3, and 4, respectively. This exercise was repeated with a range of standard errors and using the NC OLS measure as the baseline with no qualitative difference in the results. Thus, the systematic changes that we report in Table 1 cannot be attributed to random perturbation. The authors thank an anonymous referee for suggesting this comparison.

Table 2
Correlations between school effectiveness measures and various grade/school level characteristics[a]

	Average 5th grade math/read score	Percent free lunch eligible[c]	Percent black[c]
South Carolina			
Math			
5th grade avg.[b]	1.00	−0.58	−0.40
Student gain index	0.58	−0.24	−0.17
Fixed effects	0.60	−0.25	−0.18
Instrumental variables	0.41	−0.15	−0.11
F.E. & I.V.	0.41	−0.15	−0.11
Reading			
5th grade avg.[b]	1.00	−0.61	−0.42
Student gain index	0.48	−0.21	−0.16
Fixed effects	0.49	−0.22	−0.16
Instrumental variables	0.31	−0.11	−0.09
F.E. & I.V.	0.33	−0.12	−0.09
North Carolina			
Math			
5th grade avg.[b]	1.00	−0.58	−0.40
NC OLS	0.60	−0.21	−0.16
Fixed effects	0.62	−0.22	−0.16
Instrumental variables	0.44	−0.11	−0.09
F.E. & I.V.	0.44	−0.12	−0.10
Reading			
5th grade avg.[b]	1.00	−0.61	−0.42
NC OLS	0.48	−0.21	−0.16
Fixed effects	0.50	−0.22	−0.16
Instrumental variables	0.31	−0.11	−0.09
F.E. & I.V.	0.32	−0.12	−0.10

[a] Owing to computational constraints associated with estimation of the combined Fixed Effect and Instrumental Variable models, the results reported in Table 1 are based on all matched test data for a randomly chosen sub-sample of 763 elementary schools. With the exception of the two F.E. and I.V. models, all calculations were also done using the entire sample of 997 schools. There are no substantial differences between the two samples. All model definitions are as in Table 1.

[b] Fifth grade averages are for matched students only.

[c] Percent Black and Free Lunch Eligibility data are taken from the 1994/95 Common Core of Data.

with SES or race are attributable to measurement error. For example, based on the South Carolina approach for math, the correlation with percent free lunch falls from −0.240 to −0.153 and with percent black from −0.172 to −0.111. Similarly the correlations fall significantly with the correction for measurement error and specification error using the North Carolina model.

In sum, correcting the estimates for the specification error has little effect on the correlations but correcting for measurement error is important. Without that correction, schools serving low-performing students and students from disadvantaged backgrounds are viewed as being less effective than they really are and those serving high-performing students are viewed as being more effective than they really are. Correcting for measurement error does not eliminate completely the correlation between school effectiveness and the SES of the students, but it reduces it significantly.

4.5. Alternative explanations

Two possible objections to our analysis need to be addressed. The first is that the functional form for the estimating equation could be incorrect. The second is that third grade test scores could belong in the fifth grade achievement equation. Neither objection is supported by the data.

The first issue is whether the true functional form for the fifth grade equation is sufficiently non-linear to generate patterns of effective schools that would differ significantly from the ones that emerged in North and South Carolina. To examine this possibility we impose no func-

tional form on the relationship between fifth and fourth grade achievement and use nonparametric estimation techniques to generate the results in Fig. 1. Each of the panels in the figure represents the relationship between the expected value of individual fifth grade achievement (the sum of math and reading for each student) and a student's observed level of fourth grade achievement, controlling for average fourth grade achievement in the school.[26] The relationship is plotted for five levels of school-wide average fourth grade achievement. The figures indicate no clear non-linearities in the achievement relationship and, hence, suggest no clear biases from using a linear or quadratic estimating equation.

The second issue arises because certain assumptions about the underlying model could imply that third grade test scores belong in the fifth grade achievement equation (see, for example, the cumulative achievement model discussed by Boardman & Murnane, 1979). Under other assumptions, however, such as that the school indicator variables and the relevant unobserved student characteristics are not correlated after controlling for lagged achievement, the third grade test scores do not belong in the equation. Which set of assumptions is closer to reality is not clear.

To examine this issue empirically, we re-estimated the South Carolina model including third as well as fourth grade test scores as explanatory variables. Because the third grade scores were in the equation, they were no longer available as instruments and hence we were not able to correct simultaneously for measurement error. The new equations generated estimated school effects that were almost identical to those generated by the equation that excluded the third grade scores (with no correction for measurement error).[27] Based on these results, we believe that measurement error rather than this form of specification error is the more serious problem.

5. Unintended incentive effects of value-added measures

Importantly, even after correcting for the measurement error in the regression model, we still find that the measured effectiveness of a school remains correlated with the mix of students in the school and that, in particular, the higher the average socioeconomic advantage and/or test scores of the school's students, the more effective the school will appear. Fig. 2 sheds further light on the basic relationship. The panels in the figure are cross sections from the non-parametric regression, referred to above, of fifth grade individual achievement (the sum of reading and math scores) on fourth grade individual achievement and average school-wide fourth grade achievement. Each panel depicts the relationship between a student's expected fifth grade achievement and the school's average fourth grade achievement, given a fixed level of the student's fourth grade achievement.

The upward sloping lines in the first four panels imply that the expected fifth grade score—and hence the expected gain in test scores—is greater for a student in a school with higher-scoring schoolmates than one with lower scoring schoolmates. Such a relationship could reflect any or all of the explanations we mentioned earlier: schools with higher achieving students may be able to attract better teachers, they may have more resources, they may generate positive peer effects, or they may be more efficient in delivering education services. The bottom panel shows, in contrast, that the expected gain for a student who comes into fifth grade with an extremely low fourth grade score (one that is in the bottom 1 percent of the overall distribution) declines the higher the average score of his schoolmates. For such a student, any positive resource or managerial effects associated with higher performing students are apparently offset by larger negative effects.

Because there are very few students of the type depicted by the bottom panel (even in schools with low average test scores), the general pattern is one of higher value added in schools with higher average student performance. The implications of this pattern are important. It implies that value-added measures of school effectiveness of the type used by North or South Carolina exacerbate any existing incentives for teachers to avoid schools in which students are on average low performing in favor of those in which students perform better on average.

6. Discussion and conclusion

Measuring each school's value added is a worthwhile endeavor, but one fraught with challenges. Some of the challenges, such as developing a consensus on the primary mission of the schools and finding reliable and valid measures of student mastery of the curriculum, are beyond the scope of this paper. Others are technical, such as assuring the availability of student test score data that are matched by student from one year to the next. Even with such data, however, a state must proceed with care

[26] Kernel-based techniques were used to non-parametrically regress the sum of individual fifth grade math and reading scores on the sum of individual fourth grade math and reading scores and the average sum of fourth grade math and reading scores in the student's school. A normal kernel was used with a bandwith of 1.7 and a sample size of 28 218. For a further discussion of the issues associated with kernel-based techniques see Lee (1996).

[27] For example, the correlation between the value-added measures under the SC fixed effects model estimated with and without third grade scores is 97% and 95% for math and reading, respectively.

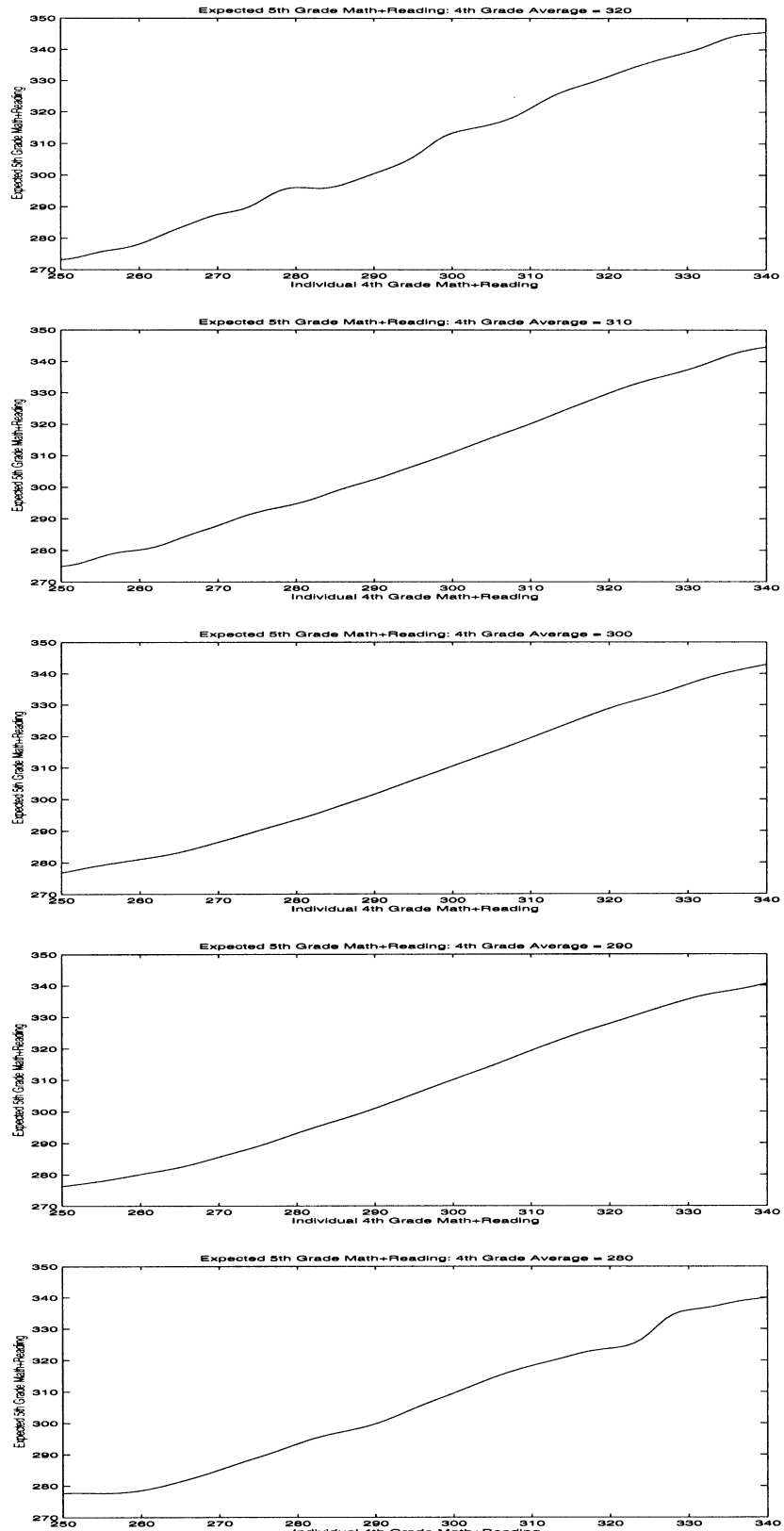

Fig. 1. Non-parametric regression: expected fifth grade achievement by fourth grade individual achievement—fourth grade class average achievement held constant.

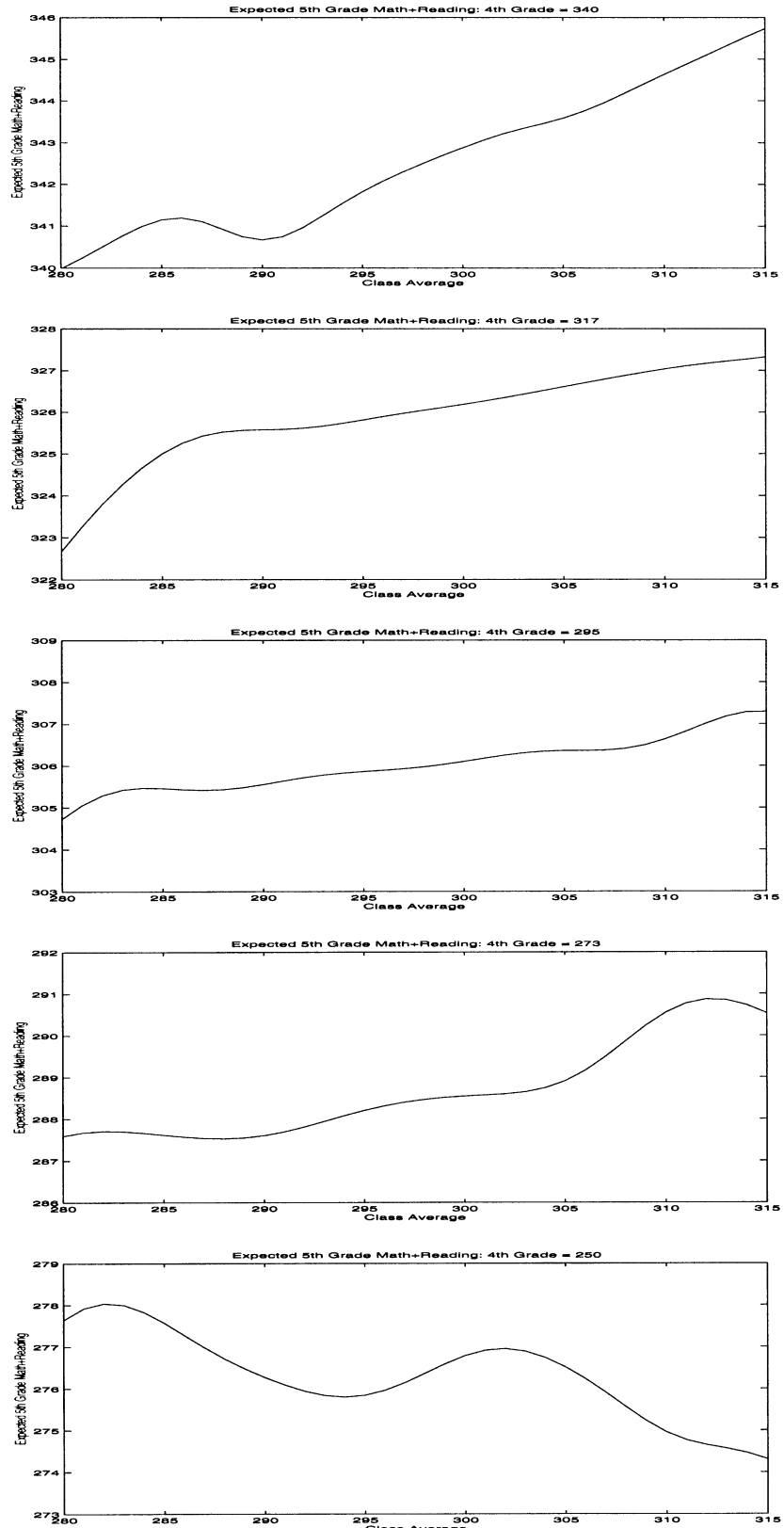

Fig. 2. Non-parametric regression: expected fifth grade achievement by fourth grade class average achievement—individual fourth grade achievement held constant.

in the development of value-added measures of school effectiveness and in their interpretation and use.

For the reasons discussed in Section 2, data limitations and other considerations typically make it infeasible for states (or districts) to implement fully specified models that would generate accurate measures of the efficiency with which schools use their resources. Instead, they implement simpler measures based exclusively on test scores. These simpler measures provide information that is useful for some, but not all, purposes. The danger is that policy makers may attribute the lackluster gain in achievement of a school's students to ineffective teachers or managers when in fact there may be other explanations such as inadequate resources or other factors outside the immediate control of the school personnel.

Thus, specification and measurement errors aside, the measures of school performance of the type used by North and South Carolina are potentially problematic for the main purpose for which they are currently being used, that is, as the basis for rewards and sanctions for school personnel. As a result, the incentives for teachers and administrators may be counterproductive in some situations. Consider, for example, schools that serve large concentrations of disadvantaged students and that do not have sufficient compensatory resources to offset the educational challenges that such students pose. In that case, schools may be deemed ineffective despite using their insufficient resources more productively and efficiently than other schools. This potential problem is exacerbated by the statistical bias against schools serving disadvantaged (low SES) students that results from failure to account for measurement error.

The combined result may well be that high quality teachers and administrators try to avoid schools serving low SES students in favor of schools serving high SES students. While anecdotal evidence from North Carolina is consistent with this view, we are not aware of any systematic study of the magnitude of this effect and believe it deserves further investigation. The larger this incentive effect, the more the accountability system would reduce the quality of education in the schools where achievement gains are most needed. However, despite these intended side effects, we emphasize that, as a measure of school effectiveness, gains in student performance are far superior to the alternative of relying on the average level of student achievement.

By focusing attention on the unintended and undesirable incentive effects of the value added measures implemented by North and South Carolina, we hope to highlight how important it is for states or districts to supplement any test-based accountability systems with other policies explicitly designed to improve the outcomes of students in schools with large concentrations of low-performing students. Otherwise, test-based accountability systems could lead to a significant widening of the gap between low-performing and high-performing students.

Acknowledgements

We thank Becky Roselius for her contributions to an earlier version of this paper. We also appreciate the constructive comments of the anonymous reviewers.

Appendix A

Tables 3 and 4

Table 3
Comparison of average test scores for sample and population (number of observations)

	Matched sample (3 years)	Unmatched data
3rd grade reading (1993)	144.12 (37 968)	142.71 (85 381)
4th grade reading (1994)	149.14 (37 968)	147.91 (85 311)
5th grade reading (1995)	153.50 (37 629)	152.36 (86 150)
third grade math (1993)	141.65 (37 852)	139.82 (85 191)
4th grade math (1994)	148.80 (37 671)	147.91 (85 311)
5th grade math (1995)	155.88 (37 611)	154.41 (86 160)

Source: NC test score data on CD-ROM, NC Department of Public Instruction.

Table 4
Race and sex components of matched and unmatched data

	Matched sample	All 3rd grade	All 4th grade	All 5th grade
% female	50.0	49.1	48.9	49.0
% white	72.7	67.0	66.1	65.7
% black	24.0	29.1	29.3	28.9

References

Altonji, J. G. (1995). The effects of high school curriculum on education and labor market outcomes. *Journal of Human Resources, 30* (3), 409–438.

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education, 52*, 113–121.

Bryk, A., & Hermanson, K. (1992). Educational indicator systems: Observations on their structure, interpretation, and use. *Review of Research in Education, 19*, 451–484.

Clotfelter, C., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. Ladd, *Holding schools accountable: performance-based reform in education*. Washington, DC: Brookings Institution.

Darling-Hammond, L. (1992). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85* (3).

Duncombe, W., Ruggiero, J., & Yinger, J. (1996). Alternative approaches to measuring the cost of education. In H. F. Ladd, *Holding schools accountable: performance-based reform in education*. Washington, DC: Brookings Institution.

Elmore, R., Abelmann, C., & Fuhrman, S. (1996). The new accountability in state education reform: From process to performance. In H. Ladd, *Holding schools accountable: performance-based reform in education*. Washington, DC: Brookings Institution.

Evans, W., Oates, W., & Schwab, R. (1992). Measuring peer group effects: A study of teenage behavior. *Journal of Political Economy, 100* (5).

Fiske, E. B., & Ladd, H. F. (2000). *When schools compete: a cautionary tale.* Washington, DC: Brookings Institution.

Greene, W. (1993). *Econometric analysis.* (2nd ed.). New York: Macmillan Publishing Company.

Hanushek, E. A., & Taylor, L. L. (1990). Alternative assessments of the performance of schools: measurement of state variations in achievement. *Journal of Human Resources, XXV* (2), 179–201.

Hausman, J. (1978). Specification and estimation of simultaneous equations models. *Econometrica, 46*.

Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek, & D. W. Jorgenson, *Improving America's schools: the role of incentives*. Washington, DC: National Academy Press.

Ladd, H. F. (1996). *Holding schools accountable: performance-based reform in education*. Washington, DC: Brookings Institution.

Lee, M. (1996). *Methods of moments and semiparametric econometrics for limited dependent variable models.* New York: Springer.

Manski, C. (1993). Identification of endogenous social effects: the reflection problem. *Review of Economic Studies, 60*.

Meyer, R. (1996). Value-added indicators of school performance. In E. A. Hanushek, & D. W. Jorgenson, *Improving America's schools: the role of incentives*. Washington, DC: National Academy Press.

North Carolina State Board of Education, 1996a. *Accountability brief.*

North Carolina State Board of Education, 1996b. *School-based management and accountability procedures.*

Sanford, E. and Thissen, D. (1995). Assessing performance across years (June 1995 revision). Processed.