

Abstract

In this paper, we analyze the statistical properties of school test scores and explore the implications for the design of school accountability systems. Using data from North Carolina, we decompose the variance of school-level test scores into a persistent component, sampling variation and other non-persistent shocks. We find that sampling variation and other non-persistent components of school test score measures account for between 25 and 65 percent of the variation in school-level test scores, with measures based on first-differences (also known as “value-added” measures) being the least reliable. We use the information from the variance decomposition to construct “filtered” estimates of the persistent component of school performance. We find that filtered estimates are more accurate indicators of school performance than conventional estimates, and forecast between 10 and 70 percent of the persistent variation in test scores one to two years into the future. Finally, we show how the optimal incentive contract would incorporate both the filtered estimate of a school’s performance as well as the conventional measure of current performance in a simple and empirically feasible manner .

Thomas J. Kane
School of Public Policy and Social Research
UCLA
3250 Public Policy Building
Los Angeles, CA 90095-1656
tomkane@ucla.edu

Douglas O. Staiger
Department of Economics
Dartmouth College
Hanover, New Hampshire 03755
douglas.o.staiger@dartmouth.edu

I. Introduction

Over the last decade, state governments have constructed elaborate school accountability systems using student test scores. By the spring of 2002, forty-three states were using student test scores to rate school performance, with twenty states attaching explicit monetary rewards or sanctions to each school's test performance (Education Week, 2002). For example, California spent \$700 million on teacher incentives in 2001, providing bonuses of up to \$25,000 to teachers in schools with the largest test score improvements. Recent federal legislation mandates even broader accountability, requiring states to test children in grades three through eight within three years, and to intervene in schools failing to achieve adequate yearly progress – initially by offering their students the choice to attend other public schools in the district, and eventually by imposing more severe sanctions such as school reorganization.

In the rush to design accountability systems, little attention has been paid to the imprecision of the test score measures on which such systems are based. The imprecision of test score measures arises from two sources. The first is sampling variation, which is a particular problem in elementary schools with an average of only 60 students per grade. A second source of imprecision arises from one-time factors that are not sensitive to the size of the sample: bad weather on the day of the test or a particularly disruptive student in a class. The volatility in test scores generated by small samples and other one-time factors can have a large impact on a school's ranking, simply because schools' test scores do not differ dramatically in the first place. This reflects the longstanding finding from the Coleman report (1966), that less than 15 percent of the variance in student test scores is between-schools rather than within-schools. Moreover, this imprecision can wreak havoc in school accountability systems: School personnel are

punished or rewarded for results beyond their control, large year-to-year swings in school rankings erode public confidence in the system, and small schools are singled out for both rewards and sanctions simply because of more sampling variation in their test score measures.

In this paper, we analyze the statistical properties of school test scores and explore the implications for the design of school accountability systems. Using six years of data on test scores from North Carolina, we apply methods developed in McClellan and Staiger (1999) to decompose the variation across schools and over time into three components: that which reflects persistent differences in performance between schools, sampling variation, and other non-persistent differences among schools. We find that sampling variation and other non-persistent components of school test score measures account for between 25 and 65 percent of the variation, with reading scores and value-added measures (based on each student's gain in test score over the previous year) being the least reliable.

The second part of the paper uses the information from the variance decomposition to construct more accurate predictions of the persistent component of school performance. In particular, we use the estimated variance components in an empirical Bayes framework to construct "filtered" estimates of each school's performance measures. The filtering technique yields estimates of the posterior mean of the persistent component of each school's performance, efficiently incorporating all available prior information on a school's test scores. We find that filtered estimates are more accurate indicators of school performance than conventional measures (such as those based upon single-year means), and forecast between 10 and 70 percent of the persistent variation in test scores one to two years into the future.

In the final part of the paper, we explore the implications our empirical results have for the design of school accountability systems. The optimal design must address three key features of test score data: noisy measures, multiple years in which to observe and reward performance, and persistent differences in performance across schools. All three of these features are addressed in a model developed by Gibbons and Murphy (1992) to study optimal incentive contracts in the presence of career concerns. This model suggests an optimal incentive contract that is closely related to our empirical results and that could be implemented in practice as part of an accountability system. Both filtered estimates of performance (which reward persistent differences in ability) and the deviation between current test scores and filtered estimates (which reward current effort) are part of an optimal incentive contract. In addition, the optimal contract places less weight on current test scores for small schools (because these schools have noisier performance measures) and in early years of an accountability system (because early test scores influence future filtered estimates, providing already strong implicit incentives).

II. The North Carolina Test Score Data

We obtained math and reading test scores for nearly 200,000 students in grades 4 and 5, attending one of roughly 1300 elementary schools each year between the 1992-93 and 1998-99 school years. The data were obtained from the N.C. Department of Public Instruction. Although the file we received had been stripped of student identification numbers, we were able to match a student's test score in one year to their test score in the previous year using data on date of birth,

race and gender.¹ In 1999, 84 percent of the sample had unique combinations of birth date, school and gender. Another 14 percent shared their birth date and gender with at most 1 other students in their school and grade and 2 percent shared their birth date with 2 other people. Less than 1 percent shared their birth date and gender with 3 or more students in the school and no match was attempted for these students. Students were matched across years only if they reported the same race. If there was more than 1 person with the same school, birth date, race and gender, we looked to see whether there were any unique matches on parental education. If there was more than one person that matched on all traits-- school, birth date, race, gender and parental education-- the matches that minimized the squared changes in student test scores were kept.

However, because of student mobility between schools or student retention, the matching process was not perfect. We were able to calculate test score gains for 66 percent of the 4th and 5th grade students in 1999. The matching rate was very similar in other years. Table 1 compares the characteristics of the matched and the non-matched sample of fourth and fifth grade students in 1999. The matched sample had slightly higher test scores (roughly .2 student level standard deviations in reading and math), a slightly higher proportion female, a slightly lower proportion black and Hispanic, and a slightly lower average parental education than the sample for which no match could be found.

¹In addition, the survey contained information on parental educational attainment reported by students. Given changes in student responses over time, we did not use parental education to match students from year to year, although we did use these data to when attempting to control for the socioeconomic background of students.

III. Volatility in Student Test Scores

In the analysis in this paper, individual student test scores and test score gains on consecutive end-of-year tests were first regression-adjusted, including dummy variables for parent's education, gender and race/ethnicity as well as fixed effects for each school and year. Figure 1 portrays the distribution of regression-adjusted mean math and reading 5th grade gain scores for each school by the number of students in each grade. (The mean gain is the average change in students' test scores between the end of grade 4 and the end of grade 5. In the economics of education literature, these are often referred to as "value-added" measures.) Test scores have been reported in standard deviations of student-level math test scores. Two facts are immediately apparent: First, virtually all of the schools with the highest mean test scores or mean gain scores (for example, schools with mean test scores more than one student-level standard deviation above the mean) were small schools, with fewer than 40 students per grade level. However, there was little difference in mean performance by school size. Indeed, the poorest performing schools were also primarily small schools. Second, this variability in test scores among small schools is not solely due to heterogeneity among small schools. The graphs on the right in Figure 1 report the change between calendar years in each school's mean test score and mean gain score. The small schools are also much more likely to report large *changes* in mean scores and mean gains from one year to the next, both positive and negative.

Table 2 provides another illustration of the volatility in school test score measures. Each year between 1994 and 1999, we ranked schools by their average test score levels and average test score gains in 5th grade, after adjusting for race, parental education and gender. We counted the proportion of times each school ranked in the top 10 percent over the six-year period. If

there were “good” and “bad” schools which could be observed with certainty, we might expect to see 90 percent of schools never ranking in top 10 percent and 10 percent of schools always ranking at the top. At the opposite extreme, where schools were equal and the top 10 percent were chosen by lottery each year, we would expect 47 percent schools ranking in the top 10 percent at least once over 6 years and only 1 in a million ranking in the top 10 percent all 6 years.

The rankings generally resemble a lottery, particularly in gain scores (value-added). If math scores were the metric, between 31 and 36 percent of schools would have ranked in the top 10 percent at some point over the 6 years, depending upon whether one used the mean test score or the mean gain in test scores. Less than 1 percent of schools ranked in the top 10 percent all 6 years. Reading test scores seem to have an even larger random component, with 35 to 38 percent of schools ranking in the top 10 percent at some point over 6 years and less than one percent of schools ranking in the top 10 percent for all 6 years. No school ranked in the top 10 percent on 5th grade reading gains for all 6 years.

Small sample size is a particularly large problem for elementary schools. However, the problem is not unique to elementary schools. Figure 2 portrays the distribution of sample sizes by grade in North Carolina. School size is generally smaller in grade 4. However, while the size of the average middle school is larger than the size of the average elementary school, there is also more heterogeneity in school size. The same phenomenon is exaggerated at grade 9. In other words, elementary schools tend to be smaller than middle schools and high schools. However, they are also more uniform in size, meaning that schools have a more similar likelihood of having an extremely high or extremely low score due to sampling variation. Middle schools and

high schools have larger sample sizes on average, but there is greater heterogeneity in the likelihood of seeing an extremely high or extremely low test score due to sampling variation.

IV. Variance Decomposition and Filtering for School Performance Data

The preceding section suggests that test scores are a very noisy measure of school performance, particularly for small schools. In this section, we develop an empirical model that decomposes the variance in school test scores into signal and noise components, and then uses this information to construct more accurate estimates of true school performance. We employ an empirical Bayes approach (Morris, 1983) similar to that developed by McClellan and Staiger (1999) to evaluate hospital performance. The estimation proceeds in two steps. The first step uses OMD methods to decompose the variance and covariance of the observed performance measures over time into signal and noise (e.g. estimation error) components, with the signal component being further decomposed into persistent (over time) and non-persistent components. The second step uses the information on the variance components to form optimal linear predictions (or forecasts) of each performance measure (or any of its components) as a function of all the observed performance measures. We refer to these predictions as “filtered” estimates because they are formally similar to estimates derived using the Kalman filter in time series (Hamilton, 1994), and because the key advantage of such estimates is that they optimally filter out the noise in the observed performance measures.

The filtered estimates have a number of attractive properties. First, they incorporate information from many performance measures and many years in a systematic way into the predictions of any one performance measure, rather than relying on ad hoc averaging across

measures or years. In particular, the filtered estimates are optimal linear predictions in the sense that they minimize the mean squared error of the prediction and thereby reduce noise in these measures to the maximum extent possible. A second property of these predictions is that, under assumptions of normality, they provide estimates of the posterior mean (and distribution) of each performance measure, conditional on all available data. As we discuss in the final section of the paper, estimates of the posterior mean are a key component of optimal incentive contracts. A third property of the filtered estimates is that regression coefficients are not attenuated when these estimates are used as independent variables (see Hyslop and Imbens, 2000). In contrast, coefficient estimates are attenuated towards zero when using conventional performance measures as independent variables because of the classical measurement error in conventional estimates. A final property of the filtered estimates is that they are quite easy to construct, given the small number of parameters to be estimated and the linearity of the filtering method. In contrast, estimation complexity has limited the application of most existing Bayesian approaches to relatively small samples (for example, Normand, Glickman, and Gatsonis, 1997).

A. *Setup and Notation*

Let $\hat{\delta}_{jt}$ be a 1x2 vector containing estimates of two different performance measures for a particular school ($j=1,\dots,J$) in a given year ($t=1,\dots,T$). Each of these estimates is derived from a student-level regression of the form:

$$(1) \quad Y_{ijt} = \delta_{jt} + X_{ijt}\beta + u_{ijt} ,$$

where Y_{ijt} would be a student performance measure (e.g. 5th grade math or reading test) and X_{ijt}

would include any relevant student characteristics that are being controlled for. The key parameters of interest are the school-specific intercepts for each year (δ_{jt}). Without loss of generality, we will assume that X includes year dummies and the school-specific intercepts are normalized so as to be mean zero in each year. Thus, with no other covariates, these would simply represent the difference between the mean test scores for each school and the average test scores for all schools in each year.

Let δ_j be the $1 \times 2T$ vector of school-specific intercepts from all years and for both performance measures. Estimation of equation (1) by standard fixed-effects methods yields unbiased estimates of these school-specific intercepts ($\hat{\delta}_j$) along with variance estimates for these parameters (S_j), where:

$$(2) \quad \hat{\delta}_j = \delta_j + \varepsilon_j$$

and S_j is an estimate of the $2T \times 2T$ variance matrix of the estimation error (ε_j). In other words, equation 2 states that school-specific estimates are composed of a signal component and a noise component, and the variance of the noise component is known.

Finally, we characterize the time-series properties of true school performance by assuming that school performance in each year (δ_{jt}) consists of two independent components: a persistent component (θ_{jt}) that is meant to capture differences across schools in curriculum, staff and facilities that would be expected to persist over time; and a non-persistent component (ψ_{jt}) that is meant to capture other idiosyncratic factors that do not disappear with sample size-- such as the weather on the day of the test or the presence of a particularly disruptive student in that

year. The persistent component is modeled as a 1st-order vector autoregression (VAR), with the two performance measures in each year depending on the previous years values of both measures plus new innovations that can be correlated across the measures. The non-persistent component is assumed to be independent across years, but allows for correlation in this component across performance measures in a given year. More specifically, we assume:

$$(3) \quad \delta_{jt} = \theta_{jt} + \psi_{jt},$$

where: ψ_{jt} is i.i.d. with $\text{Var}(\psi_{jt}) = \Psi$

and: $\theta_{jt} = \theta_{j,t-1}\Phi + v_{jt}$, where v_{jt} is i.i.d. with $\text{Var}(v_{jt}) = \Sigma$, and initially $\text{Var}(\theta_{j1}) = \Gamma$.

The unknown parameters of this model are the variances and covariance of the non-persistent component (Ψ), the variances and covariance of the persistent component in the first year of the data (Γ) and the innovations to the persistent component (Σ), and the coefficients determining how lagged values of each performance measure influence current values (Φ).

B. Estimation

The ultimate goal is to develop linear predictions of each school's performance (δ_j) as a function of the observed performance estimates ($\hat{\delta}_j$). In general, the minimum mean squared error linear predictor is given by $\hat{\delta}_j \beta_j$ where $\beta_j = [E(\hat{\delta}_j' \hat{\delta}_j)]^{-1} E(\hat{\delta}_j' \delta_j)$. Moreover, if we assume normality in equations 2 and 3, then this estimator also gives the posterior mean ($E[\delta_j | \hat{\delta}_j]$) and is the optimal choice for any symmetric loss function. But this estimator depends

upon two unknown moment matrices -- $E(\hat{\delta}_j' \hat{\delta}_j)$ and $E(\hat{\delta}_j' \delta_j)$ -- which, based on equation 2,

can be further decomposed as follows:

$$(4) \quad E(\hat{\delta}_j' \hat{\delta}_j) = E(\delta_j' \delta_j) + E(\varepsilon_j' \varepsilon_j)$$

$$(5) \quad E(\hat{\delta}_j' \delta_j) = E(\delta_j' \delta_j)$$

Thus, constructing the optimal linear predictor requires estimates of the signal variance $[E(\delta_j' \delta_j)]$ and the noise variance $[E(\varepsilon_j' \varepsilon_j)]$ for each school. Similarly, to construct predictions of the persistent component in school performance (θ_j) requires estimates of the variance in the persistent component $[E(\theta_j' \theta_j)]$. Therefore, the first step in constructing the optimal linear predictor is estimating each of the variance components.

We estimate each of the variance components as follows. As mentioned above, estimation of equation 1 with the individual student test score data generates estimates of $[E(\varepsilon_j' \varepsilon_j)]$ for each school-- namely S_j , the variance-covariance matrix for the school-specific intercepts. The variance in the signal and in the persistent component can each be calculated as a function of the parameters of the time-series model specified in equation 3, e.g. $[E(\delta_j' \delta_j)] = f(\Psi, \Phi, \Sigma, \Gamma)$. We estimate the time series parameters (Ψ , Σ , Γ and Φ , all 2x2 matrices, the first three symmetric) by noting that equation 4 implies that:

$$(6) \quad E\left[\hat{\delta}_j' \hat{\delta}_j - S_j\right] = E\left[\delta_j' \delta_j\right] = f(\Psi, \Phi, \Sigma, \Gamma)$$

Thus, the time-series parameters can be estimated by Optimum Minimum Distance

(OMD) methods (Chamberlain, 1983), i.e. by choosing the parameters so that the theoretical moment matrix, $f(\Psi, \Phi, \Sigma, \Gamma)$, is as close as possible to the corresponding sample moments from the sample average of $\hat{\delta}_j' \hat{\delta}_j - S_j$. Let d_j be a vector of non-redundant (lower triangular)

elements of $\hat{\delta}_j' \hat{\delta}_j - S_j$, and let $g(\Psi, \Phi, \Sigma, \Gamma)$ be a vector of the corresponding moments from the theoretical moment matrix. The OMD estimates of $(\Psi, \Phi, \Sigma, \Gamma)$ minimize the objective function:

$$(7) \quad q = N \left[\bar{d} - g(\Psi, \Phi, \Sigma, \Gamma) \right]' V^{-1} \left[\bar{d} - g(\Psi, \Phi, \Sigma, \Gamma) \right]$$

where V is the sample covariance for d_j , and \bar{d} is the sample mean of d_j . The value of the objective function (q) provides a measure of goodness of fit, and is distributed $\chi^2(p)$ if the model is correct, where p is the degree of over-identification (the difference between the number of elements in d and the number of parameters being estimated).

V. Results

A. *Decomposing the Variation in 5th Grade Math and Reading Scores.*

Table 3 reports the parameter estimates for fifth grade math and reading scores, both levels and gains, after having adjusted each student's level and gain score for race, gender and categories of reported parental education (using the same six categories reported in Table 1).

There are 4 findings worth noting in Table 3: First, after sorting out differences across schools that are due to sampling variation and other non-persistent differences in performance,

the variation in school performance is small relative to the differences in student performance. Our estimate of the variance in adjusted school math test scores in 1994 (the variance in initial conditions, Γ) was just .061, implying a standard deviation of .247. Since test scores have been reported in student-level standard deviation units, this implies that the standard deviation in underlying differences in math performance at the school level is only one quarter of the standard deviation in 5th grade math test scores at the student level. There is even less of a signal to be found in reading test scores and in gain scores for both reading and math. Indeed, we estimate that the standard deviation in school reading score gains in fifth grade is only .077, one-thirteenth as large as the standard deviation in reading test scores among students.

Second, schools that perform well in 5th grade math also tend to perform well in 5th grade reading, in both levels and gains. This fact is reflected in the high correlations between reading and math in the initial conditions and persistent innovations. Even the non-persistent innovations are highly correlated for reading and math. This may not be surprising, given that elementary school teachers tend to teach both reading and math subjects. A disruptive student in the class or a particularly effective teacher would affect both math and reading test scores for a particular school. As a result, any differences in instructional quality are likely to be reflected in both.

Third, while the signal variance shrinks when moving from test score levels to gains-- as we might expect, as long as some of the differences in test score levels reflect the different starting points of students in the schools-- the variance in the non-persistent innovations in school performance does not. In fact, the non-persistent variance in math test scores doubles when moving from levels to gains, from .006 to .011. Recall that the variance in non-persistent

innovations (Ψ) reflects the changes in student test scores that do not disappear with the number of students in the school, yet do not seem to be passed on from one year to the next. Some sources of such variation would be factors that would affect a whole school-- such as a dog barking in the parking lot on the day of the exam, a rainy day, an active construction site next door or transient factors affecting a single classroom-- one or two particularly disruptive students or strong “chemistry” between the teacher and the particular group of students. Because the gain scores use the difference in outcomes over two testing dates, such non-persistent variation would tend to be magnified (because there is no covariance between them). Meanwhile, it is the ratio of the variance in the signal to the total variance in school test scores that determines the degree to which the raw data are misleading. That ratio falls when moving from test score levels to gain scores, both because the variance of the persistent signal shrinks and because the non-persistent variation rises.

Fourth, while we estimate that there is a considerable amount of persistence in math and reading performance over time at the school level (both levels and gains), we can strongly reject the hypothesis that school test scores are fixed effects. If school rankings were fixed over the long term, albeit with some short term fluctuations due to sampling variation and other non-persistent changes in school performance, we would expect the coefficients on lagged scores to approach one and for the variance in the persistent innovations to approach zero. Aside from sampling variation, the only innovations would be of the non-persistent type. As reported in Table 3, the coefficients on the lagged values range from .7 to .98, but generally range from .7 to .8. Moreover, the variance in persistent innovations-- changes in school performance that are at least partially passed on from one year to the next-- is estimated to be positive for all four

outcomes. The p-values on the test of the fixed-effect specification reported in Table 3 suggest that the fixed effect specification can be rather comfortably rejected.

Interestingly, the degree of persistence in school mean gain scores over time roughly matches the proportion of a school's teachers who remain from one year to the next. Figure 3 reports the implied persistence in math and reading mean gain scores implied by the estimates in Table 3, as well as the proportion of elementary teachers in North Carolina who were employed in the same school for 1 through 5 years or more. The latter estimates are derived from 908 respondents to the 1993-94 Schools and Staffing Survey who reported teaching in North Carolina.² For instance, 85 percent of teachers are estimated to remain in the same school for 1 or more years, 75 percent for 2 or more years, 67 percent for 3 or more years, 57 percent for 4 or more years and 50 percent for 5 or more years. The correlation in a school's math and reading value-added fades only slightly more rapidly. For instance, after 5 years, the correlation in 5th grade math gains was roughly .3 while the correlation in 5th grade reading gains was roughly .35. While this evidence is far from conclusive, it suggests that one reason for the lack of strong correlation in school performance over time may be the high rate of teacher turnover.³

Table 4 uses these estimates to decompose the variance in school mean test scores for students in the 5th grade in 1994 into 3 parts: that reflecting differences in performance that will at least partially be passed on from one year to the next (Γ), that reflecting sampling variation and

²We thank Terry Moe of the Hoover Institution at Stanford University for generously providing us with these estimates.

³Our data provide other evidence of the importance of teacher heterogeneity. For instance, the correlation in math and reading value-added is considerably higher in fourth and fifth grades, when students typically have the same instructor for all subjects than in seventh and eighth grades, when students have different instructors for each subject.

that reflecting non-persistent variation that is independent from year to year (Ψ). Because the variance due to sampling variation is sensitive to the number of 5th grade students in a particular school, we report the proportions attributable to each source of variation by school size.

In both levels and gains, reading scores contain a lower ratio of signal to noise. For small schools (25 students per grade level), only 49 percent of the variance in reading levels and only 26 percent of the variance in reading gains reflect differences that will be partially passed on from one year to the next. This is due to the fact reported above, that there is little underlying variance in schools' performance in reading after adjusting for demographic differences between the students, at least relative to math performance. The remainder is due to either sampling variation or other types of non-persistent fluctuations in scores.

Moreover, for both reading and math, gain scores contain a much larger share of noise than test score levels. Sampling variation is not the reason. Rather, it is the rising importance of other types of non-persistent variation not sensitive to sample size that contributes to the noise in gain scores. Among small schools, the proportion of variance in reading test scores due to sampling variation increases only slightly from 44 to 51 percent between reading levels and reading gains. Among these same small schools, the proportion of variance in math performance due to sampling variation actually declines from 26 to 23 percent in moving from levels to gains. However, the proportion of variance due to other types of non-persistent factors nearly triples from 8 percent to 22 percent when moving from reading levels to gains and more than triples from 7 percent to 25 percent in moving from math levels to math gains. Even in large elementary schools, with 100 students per grade level, less than half of the variance in reading gains (43 percent) and only about two-thirds of the variance in math gains (63 percent) is due to

signal, with much of the remainder due to non-persistent fluctuations other than sampling error.

B. Predicting Performance Based upon Filtered Estimates

In focusing on a school's test score performance, parents and policymakers are primarily interested in using that evidence to draw some inferences about the state of educational quality in the current academic year and in future years. Thus, the primary interest is in estimating the persistent component of school test scores, rather than the transitory components that reflect sampling error and other one-time factors affecting test performance. Assuming that our statistical model is correct, filtered estimates of the persistent component are the best linear predictor. In this section, we evaluate whether filtered estimates in fact outperform other measures in terms of estimating differences in school test scores that persist into the future.

Table 5 compares the mean performance in 1999 for schools ranking in the top 10 percent in 5th grade math gains on two different measures: the simple means of math gains in 1997 and the filtered prediction that would have been made of a school's performance in 1999 using all of the data available through 1997. Thus, both predictions use only the data from 1997 or before. However, the filtered prediction incorporates information from reading scores and from prior years, and "reins in" the prediction according to the amount of sampling variation and non-persistent fluctuations in the data.

Table 5 reports the mean 1999 performance, cross-tabulated by whether or not the school was in the top 10 percent using the filtering technique and using the naive estimate based upon the actual 1997 scores. Sixty-five schools were identified as being in the top 10 percent as of 1997 using both the naive and the filtered predictions, and these schools scored .15 student level

standard deviations higher than the mean school two years later in 1999. However, among the schools where the two methods disagreed, there were large differences in performance. For instance, among the 25 schools that the filtering method identified as being in the top 10 percent that were not in the top 10 percent on the 1997 actual scores, the average performance on 5th grade math gains was .124 student-level standard deviations above the average in 1999. On the other hand, among the 25 schools chosen using actual 1997 scores who were not chosen using the filtering technique, scores were .022 standard deviations *lower* than the average school in 1999. The next to last column and row in Table 5 reports the difference in mean scores moving across the first two columns or first two rows. Among those who were not identified as being in the top 10 percent by the filtering method, knowing that they were in the top 10 percent on the actual 1997 score provided very little information regarding test scores. In fact the test scores were -.006 standard deviations lower on average holding the filtered prediction constant. In contrast, among those were not identified as being in the top 10 percent on actual 1997 scores, knowing that they were selected using the filtering method was associated with a .140 standard deviation difference in performance. Apparently, the filtering method was much more successful in picking schools that were likely to perform well in 1999.

Moreover, the filtering technique provides a much more realistic expectation of the magnitude of the performance differences to expect. As reported in the last column of Table 5, the schools in the top 10 percent on the actual test in 1997 scored .453 standard deviations higher than the average school in 1997. If we had naively expected them to continue that performance, we would have been quite disappointed, since the actual difference in performance was only .115 standard deviations. On the other hand, among those who were chosen using the filtering

method, we would have predicted that they would have scored .180 standard deviations higher than the average school in 1999 based upon their performance prior to 1998. The actual difference in performance for these schools was .160 standard deviations.

Table 6 compares the R^2 one would have obtained using 3 different methods to predict the 1998 and 1999 test scores of schools using only the information available prior to 1998. The first method is the “filtering method” described in the methodology section above. The second method is using the actual 1997 score and applying a coefficient of unity to it when predicting the 1998 and 1999 scores. The third method would be to use the 4-year average of math performance prior to 1998 (1994-1997) to predict 1998 and 1999.

Whether one is trying to anticipate math or reading, levels or gains in 5th grade, the filtering method leads to greater accuracy in prediction. The R^2 in predicting 5th grade math levels was .41 using the filtering method, .19 using the 1997 score and .29 using the 1994-97 average. The filtering method also calculates a weighted average using the 1994-97 scores, but it adjusts the weights according to sample size (attaching a larger weight to more recent scores for large schools) and uses both the math and reading score histories in predicting either. In so doing, it does much better than a simple average of test scores over 1994-97.

In predicting math or reading gain scores in 1998, the second column reports *negative* R^2 when using the 1997 scores alone. A negative R^2 implies that one would have had less squared error in prediction by completely ignoring the individual scores from 1997 score and simply predicting that performance in every school would be equal to the state average in 1998 scores. Of course, one could probably do even better by not ignoring the 1997 score, but simply applying a coefficient of less than 1 to the 1997 score in predicting future scores. That is essentially what

the filtering method does, while recognizing that the optimal coefficient on the 1997 score (and even earlier scores) will depend upon the amount of non-persistent noise in the indicator as well as the school size.

Although it performs better than either the 1997 score or the 1994-97 average in predicting 1998 and 1999 gains, the R^2 using the filtering method is only .16 on math gains and .04 on reading gains. This hardly seems to be cause for much celebration, until one realizes that even if the filtering method were completely accurate in predicting the persistent portion of school test scores, the R^2 would be less than 1 simply because a large share of the variation in school performance is due to sampling variation or other non-persistent types of variation. Because of these entirely unpredictable types of error, the highest R^2 one could have hoped for would have been .75 in predicting math levels, .60 in predicting reading levels, .55 for math gains and .35 in reading gains. For math gains, for instance, the filtering method was able to predict 16 percentage points of the 55 percentage points that one ever had a hope of predicting, implying an R^2 for the systematic portion of school test scores of $.16/.55=.29$. Using this standard, the results in Table 6 imply that filtered estimates can forecast approximately 10 to 30 percent of the persistent variation in test score gains, and 35 to 65 percent of the persistent variation in test score levels.

VI. Optimal Incentive Contracts

In this section, we explore the implications our empirical results have for the design of school accountability systems. One important goal of a school accountability system is to give schools proper incentives to improve student performance. Thus, the design of a school

accountability system is closely related to the literature on optimal incentive contracts. The optimal incentive design must address three key features of test score data: noisy measures, multiple years in which to observe and reward performance, and persistent differences in performance across schools. All three of these features are addressed in a model developed by Gibbons and Murphy (1992) to study optimal incentive contracts in the presence of career concerns. Moreover, this model suggests an optimal incentive contract that is closely related to our empirical results and that could be implemented in practice as part of an accountability system.

Gibbons and Murphy (1992) consider a T-period model in which a worker's output in each period (y_t) is the sum of the worker's ability (η), effort (a_t), and noise (e_t). Only output is directly observable each period, while the distributions of ability (which does not change over time) and noise (which is i.i.d. over time) are normal with known mean and variance. Workers are assumed to be risk averse with convex costs of effort, so that the optimal contract will balance the stronger incentives of pay-for-performance contracts against the increased risk they impose on the worker.

Information is assumed to be imperfect but symmetric, so that both the worker and all potential firms will observe each worker's output over time and update their beliefs about ability. If one assumes that labor markets are competitive and that contracts must be renegotiation-proof, then this implies that the optimal contract must be a series of short-term contracts that earn zero expected profits for the firm in each period. Gibbons and Murphy consider compensation contracts that are linear in current output, where the intercept and slope of the compensation contract at time t may depend on information available prior to time t .

The optimal contract in this model takes a very simple form in which the wage paid in each period (w_t) is made up of two parts: a base wage (c_t) that is equal to the expected value at time $t-1$ of a worker's output in time t ; and an incentive payment that is proportional to the difference between the worker's actual and expected output in time t . The base wage is the sum of the worker's anticipated effort in period t (a_t^*) which is known in equilibrium, and the expected value of the worker's ability conditional on all information available through time $t-1$ (m_{t-1}). Since effort is known in equilibrium, the difference between a worker's output and effort ($y_t - a_t^*$) in each period provides a noisy signal of ability that can be combined with the prior distribution on ability to form the posterior mean $m_{t-1} = E[\eta | y_1 - a_1^*, \dots, y_{t-1} - a_{t-1}^*]$. Thus, the optimal contract is of the form:

$$(8) \quad w_t = c_t + b_t^* (y_t - c_t), \quad \text{and} \quad c_t = a_t^* + m_{t-1}.$$

where b_t^* is the slope of the optimal contract. Because the difference between actual and expected output is mean zero by definition, these contracts will yield zero expected profits in each period (i.e. pay workers c_t , their expected productivity).

Gibbons and Murphy derive a number of important properties of optimal incentive contracts in this model. First, worker's will have career concerns – concerns that current output will influence future base wages because the market uses this information to form beliefs about the worker's ability in the future. These career concerns will generate strong implicit incentives for workers, even in the absence of explicit contractual incentives (e.g. $b_t^* = 0$). Moreover, these implicit career concern incentives will be stronger (1) early in the career, when the market's prior beliefs about ability are least precise and when the worker has a long remaining career, and (2) when variance in ability (η) is larger or when variance in the noise (e_t) of the output measure is

smaller, both of which lead the market to place more weight on past output in setting base wages. To balance overall incentives, the optimal incentive contract reduces explicit incentives (b_t^*) when implicit career concern incentives are strongest. Thus, the reward for contemporary performance should grow with worker experience ($b_{t+1}^* > b_t^*$), and decline with heterogeneity in ability.

Finally, an increase in the noise of the performance measure has two offsetting effects: implicit career concern incentives are reduced, but explicit incentives involve more risk. On net, the second (conventional) effect dominates, particularly late in the career, and a noisier performance measure will be associated with weaker explicit incentives (lower b_t^*).

There is a natural mapping between the Gibbons and Murphy model, and therefore the implications of that model, and the statistical model of school-level test scores presented in equations (2) and (3). Consider the school administrator as the worker, with school-level test scores (or gains) being the noisy output measure reflecting the sum of ability, effort and noise. The noise component is represented in our statistical model as the sum of measurement error (ε_{jt}) and other non-persistent factors (ψ_{jt}) affecting test scores, while the ability of the school administrator is represented in our statistical model as the persistent component of school test scores (θ_{jt}). We allow for a more general stochastic specification than do Gibbons and Murphy (e.g. ability is autoregressive rather than fixed, and the noise is heteroskedastic) but this does not change any of the fundamental implications of the model. Finally, the effort exerted by the administrator to improve student performance (including effort to provide teachers with appropriate incentives) would be represented in our statistical model by a common time-period intercept if one assumes that all schools are identical except for ability (and therefore exert

identical effort in equilibrium in the Gibbons and Murphy model). In fact, equilibrium effort in the Gibbons and Murphy model depends on the noise variance, suggesting that one should group schools according to the amount of noise variance in their test scores (e.g. by size class) and allow for separate time-period intercepts to estimate effort for each group.

Based on this mapping between the theoretical and statistical model, the optimal incentive contract from equation (8) can be restated in terms of our statistical model as:

$$(8') \quad w_{jt} = \lambda \left\{ \hat{a}_t + \hat{m}_{j,t-1} + b_t^* (\hat{\delta}_{jt} - \hat{m}_{j,t-1}) \right\},$$

where λ is the marginal dollar value of an increase in average test scores, $\hat{\delta}_{jt}$ is the estimate of

school performance for school j in year t , $\hat{m}_{j,t-1} = E(\theta_{jt} | \hat{\delta}_{j1}, \dots, \hat{\delta}_{j,t-1})$ is the filtered estimate

(as described above) of school performance in period t conditional on school test-scores observed

through period $t-1$, and \hat{a}_t is the average test score across all schools in year t . In other words,

equation 8' says that the optimal incentive contract sets wages proportional to the sum of (1)

average test scores in the state (or in a peer group of similar schools), (2) the filtered estimate of

school performance this year (relative to other schools) conditional on all past performance, and

(3) a term that is proportional to the difference between actual performance this year and the

filtered estimate. Thus, both filtered estimates of performance (which reward persistent

differences in ability) and the deviation between actual performance and filtered estimates (which

reward current effort) play a role in optimal incentive contracts. Moreover, with the exception of

b_t^* , all three terms entering the optimal incentive contract can be directly estimated from school

test score data using our statistical model.

While the optimal strength of explicit incentives (b_t^*) cannot be quantitatively estimated from the data, the Gibbons and Murphy model has two qualitative implications that are of practical importance in designing optimal incentives for school administrators. First, explicit incentives should become stronger as one accumulates more information on an administrator's ability. Thus, explicit incentives can be waived early in an administrator's career (or during the first few years of a school accountability program) because implicit incentives from career concerns assure adequate effort. In other words, performance measures from early in an administrator's career would be used primarily to determine the base wage later in the career (e.g. to sort administrators into pay grades), while later in the career an administrator's performance (relative to what was expected for someone in that pay grade) would be linked to explicit financial rewards.

A second implication of the Gibbons and Murphy model (and incentive models more generally) is that small schools with noisier test score measures should have weaker explicit incentives relative to large schools, i.e. small schools should receive smaller financial rewards for similar observed performance levels (or equivalently, be required to achieve a higher level of performance to receive a similar financial award). One practical method of accomplishing this is to group schools into size classes (as is done in high school sports), and give equal financial rewards based on each school's percentile rank within their size class. For groups with more variable test scores, a one percentile change in rank will be associated with a larger change in test scores, thus requiring larger changes in performance to receive similar changes in financial rewards (e.g. weaker explicit incentives).

We have focused in this paper on measures of performance and incentive systems at the school level, rather than individual teacher level. The primary reason is that our data did not allow us to track students matched to individual teachers over time. Moreover, most states have constructed their existing accountability systems to focus at the school, rather than teacher level. There may be legitimate reasons for this. For instance, principals may have access to data in addition to student test scores, such as direct classroom observation, which we would encourage them to employ with school-level incentives. We would also want to encourage teachers in a given school to cooperate. Moreover, the teacher-level test score measures may be even more noisy than the school level measures, given the small sample sizes for individual teachers. Nevertheless, with the appropriate data, all of the above measures could be translated to the teacher level.

VII. Conclusion

School-level test scores are the foundation upon which states are building their school reform strategies. Our results suggest that sampling variation and other non-persistent components of school test score measures account for a large fraction of the variation in school test scores, particularly for value added measures (based on test score gains). While the results in this paper focus on 5th grade scores in North Carolina, similar issues are evident in other states and other grades (Kane and Staiger (2001, 2002), Kane, Staiger and Geppert (2002)). In general, while there are important underlying differences in school performance, these are largely obscured by the noise in single-year test score measures, particularly in smaller schools.

Unfortunately, school accountability systems have been designed with little recognition of the measurement properties of school-level test scores. For example, most states have designed their accountability systems to focus on schools at the extremes. If test scores were more precise measures, it might seem reasonable to start with the schools in the upper tail of the distribution-- with the highest test scores, or the greatest improvement in test scores-- in the search for exemplars. Similarly, with a limited set of resources for reconstituting underperforming schools, it might seem reasonable to start at the lower extreme of the test score distribution. However, given the importance of sampling variation highlighted in this paper, the tails of the distribution are disproportionately populated by small schools. For example, even though there was no systematic relationship between the mean value-added measure the state uses to rate schools and school size, the smallest fifth of schools in North Carolina were more than 20 times more likely than the largest fifth of schools to be recognized among the top 25 schools in the state in 1997 and 1999 (Kane and Staiger, 2002). Moreover, all but one of the schools in North Carolina which were required to work with state assistance teams because of poor test performance were smaller than the median size. Similarly, among eligible schools in California competing to win bonuses of \$5,000, \$10,000 and \$25,000 per teacher based upon the improvement in their mean test scores between 1999 and 2000, the smallest fifth of schools were nearly four times as likely to win the bonuses as the largest fifth of schools. Ironically, the large urban schools that are the focus of much of reformers' concern may face little marginal incentive from rewards and punishments at the extremes, since even large changes in their expected performance may have little impact on the likelihood that they will be among the outliers.

Given the data provided by states on school report cards, it would be very difficult for parents to do their own adjustment for the signal and noise in school-level test scores. Some key piece of information is usually missing: for instance, states often fail to report the student-level variance in scores with which one could infer the within-school variance (in fact, in California, one is required to buy that information from the test publisher) or their web sites make it difficult to plot the volatility in school-level test scores for more than one school at a time. As a result, the inferences that are drawn from such data are often faulty. For example, when the 1998-99 MCAS test scores were released in Massachusetts in November of 1999, the Provincetown district showed the greatest improvement over the previous year. The *Boston Globe* published an extensive story describing the various ways in which Provincetown had changed educational strategies between 1998 and 1999, interviewing the high school principal and several teachers. (Tarcy, 1999) As it turned out, they had changed a few policies at the school-- decisions that seemed to be validated by the improvement in performance. One had to dig a bit deeper to note that the Provincetown high school had only 26 students taking the test in 10th grade-- the type of small sample that is likely to yield large swings in performance.

But even at large schools, volatility of test scores can be problematic. For example, the initial versions of the Elementary and Secondary Education Act (ESEA) passed by the House and Senate in 2001 required schools to show at least a one percentage point increase in math and reading proficiency *every year* or be subject to a variety of sanctions. This provision had to be rewritten in conference committee after preliminary analysis suggested that virtually every school in the country would have failed to improve at least once in a five year period because of the natural volatility in test score measures (Broder, 2001; Kane, Staiger and Geppert, 2001, 2002).

Our results have two practical implications for improving school accountability systems. The first implication is that school accountability systems could rely on much better indicators of school performance, such as the filtered estimates we have proposed. States have made arbitrary decisions to hold schools accountable to the mean for a single year or, in a handful of states, in two years. Rather than imposing an arbitrary choice of the period over which a mean is to be calculated, the filtering method proposed in this paper would systematically pool all available information from other measures and prior years in a manner which minimizes mean squared error. Moreover, the weights applied to different years will implicitly depend upon the size of the school, the estimated signal variance and the degree of persistence in school-level differences. As a result, the filtered measures provide more accurate forecasts of future school performance and are less volatile over time. Both of these features are likely to improve public confidence in the measures on which any accountability system is based.

The second implication of our results is that school accountability systems could be better designed in terms of the incentives they provide to school administrators. In the context of a model that allows for unobserved effort, heterogeneous ability, and multiple periods to observe and reward effort, the optimal contract takes a form that is both simple to implement and familiar in terms of contracts seen in other employment settings. A practical approximation of the optimal contract would have three key features. First, as is done in high school sports, it would sort schools into separate size classes in order to account for the fact that smaller schools have less reliable performance measures. Second, it would use filtered estimates of relative performance within each size class to sort schools into “pay grades,” gradually rewarding those schools that have persistently high test scores as this becomes apparent over time. Finally, once a

school's (or administrator's) performance had been observed for a sufficient number of years – e.g. after a tenure of, say, 5 years – the optimal contract would begin offering an incentive payment tied to each school's ranking within their size class based on current year test scores.

In order to focus on the statistical properties of school test scores and their implications, we have abstracted from a number of important issues. Other concerns have been raised regarding test-based accountability systems, such as whether first-differences for individual students measure the value-added by school personnel or other pre-existing differences in student trajectories, whether test-based accountability systems encourage teachers to narrow the curriculum or spend too much time on test-taking skills, or whether high stakes testing encourages teachers and students to cheat. Inquiry into each of these other issues should continue.⁴ However, even if some resolution to these psychometric challenges can be found, we would still be left with the problem of summarizing differences in school performance with imperfect measures and constructing an accountability system based upon these measures. Our goal has been to make progress on this latter problem.

⁴See Koretz (forthcoming) for an excellent review of this literature.

References:

- Broder, David S. "Long Road to Reform: Negotiators Forge Education Legislation" Washington Post, December 17, 2001, p. A01.
- Chamberlain, Gary, "Panel Data" Chapter 22 in Zvi Griliches and Michael D. Intriligator (eds.) Handbook of Econometrics, Vol. II (New York: Elsevier Science, 1984). pp. 1247-1318.
- Coleman, James S., E.Q. Campbell, C.J. Hopson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York Equality of Educational Opportunity (Washington, DC: U.S. Department of Health, Education and Welfare, 1966).
- Gibbons, Robert and Kevin J. Murphy, "Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence" Journal of Political Economy (1992) Vol. 100, No. 3, pp. 468-505.
- Hamilton, James "State-Space Models," Chapter 50 in R.F. Engle and D.L. McFadden (eds.) Handbook of Econometrics, Vol. IV, (New York: Elsevier Science, 1994), pp. 3039-3080.
- Hyslop, Dean and Guido W. Imbens "Bias from Classical and Other Forms of Measurement Error" Journal of Business and Economic Statistics (2001) Vol. 19, No. 4, pp. 475-481.
- Kane, Thomas J. and Douglas O. Staiger, "Improving School Accountability Measures," National Bureau of Economic Research Working Paper No. 8156, March 2001.
- Kane, Thomas J. and Douglas O. Staiger, "Volatility in School Test Scores: Implications for Test-Based Accountability Systems" Diane Ravitch (ed.) Brookings Papers on Education Policy, 2002, (Washington, DC: Brookings Institution, 2002).
- Kane, Thomas J., Douglas O. Staiger and Jeffrey Geppert "An Assessment of the House and Senate Education Bills" Unpublished paper, July 2001.
- Kane, Thomas J., Douglas O. Staiger and Jeffrey Geppert "Randomly Accountable: Test Scores and Volatility" Education Next, (Spring 2002) Vol. 2, No. 1, pp. 56-61.
- Koretz, Daniel "Limitations in the Use of Achievement Tests as Measures of Educator's Productivity" Journal of Human Resources (forthcoming).
- McClellan, Mark and Douglas Staiger, "The Quality of Health Care Providers" National Bureau of Economic Research Working Paper No. 7327, August 1999.

Morris, Carl. "Parametric Empirical Bayes Inference: Theory and Applications" Journal of the American Statistical Association, (1983) Volume 381, No. 78, pp. 47-55.

Normand, Sharon-Lise, Mark Glickman and Constantine Gastonis, "Statistical Methods for Profiling Providers of Medical Care: Issues and Applications" Journal of the American Statistical Association (1997) Vol. 92, No. 439, pp. 803-814.

"Quality Counts 2002: Building Blocks for Success", Education Week, (January 10, 2002) Vol. 21, No. 17, p. 75-76.

Rivkin, Steven, Eric Hanushek and John Kain, "Teachers, Schools and Academic Achievement" National Bureau of Economic Research Working Paper No. 6691, August 1998.
(Revised, April 2000)

Tarcy, Brian "Town's Scores the Most Improved" Boston Globe, December 8, 1999, p. C2.

Table 1.
Characteristics of the Matched and Non-Matched
Sample of 4th and 5th Grade Students in 1999

	Non-Matched	Matched
% of 4th and 5th Grade Students	34.2	65.8
Mean Math Score	153.8	156.5
S.D. in Math Score	11.1	10.5
Mean Reading Score	150.5	152.4
S.D. in Reading Score	9.5	9.1
Percent Female	47.4%	50.1%
Percent Black	35.1	27.7
Percent Hispanic	5.4	2.2
Parental Education:		
H.S. Dropout	16.6%	9.8%
H.S. Graduate	47.1	43.7
Trade/Business School	4.6	5.3
Community College	11.3	14.2
Four-Year College	16.5	21.9
Graduate School	3.9	5.1
Sample Size	69,388	133,305

Note: Each of the differences above were statistically significant at the .05 level.

Table 2.
Proportion Ranking in the Top 10 Percent
on 5th Grade Test Scores 1994-1999

Number of Years in Top 10% during 1994-99	Adjusted Levels		Adjusted Gains		<i>Expected Proportion</i>	
	Math	Reading	Math	Reading	<i>Annual Lottery</i>	<i>Certainty</i>
Never	.6868	.6499	.6398	.6152	.5314	.9000
1 Year	.1633	.2181	.2237	.2383	.3543	0
2 Years	.0749	.0727	.0694	.0940	.0984	0
3 Years	.0369	.0235	.0380	.0336	.0146	0
4 Years	.0190	.0179	.0213	.0179	.0012	0
5 Years	.0101	.0089	.0045	.0011	.0005	
All 6 Years	.0089	.0089	.0034	0	.000001	.1000

Note: Test scores were adjusted for the race, parental education and gender of the students and then averaged by grade level within schools.

Table 3.
Estimates of Parameters Describing Time Series of School Effects
for 5th Grade Math and Reading

	Adjusted Levels		Adjusted Gains	
	Math	Reading	Math	Reading
Coeff on Math _{t-1} Φ_1	.694 (.028)	-.075 (.020)	.767 (.046)	.006 (.025)
Coeff on Read _{t-1} Φ_2	.255 (.039)	.989 (.031)	.046 (.102)	.780 (.070)
Variance in Initial Conditions (Γ) [Implied Standard Dev.]	.061 (.004) [.247]	.030 (.002) [.174]	.023 (.002) [.152]	.006 (.001) [.077]
Correlation in Initial Conditions		.796 (.018)		.737 (.051)
Variance in Persistent Innovations (Σ) [Implied Standard Dev.]	.018 (.002) [.135]	.005 (.001) [.072]	.010 (.001) [.100]	.0016 (.0005) [.040]
Correlation in Persistent Innovations		.673 (.049)		.623 (.104)
Variance in Non-persistent Innovations (Ψ) [Implied Standard Dev.]	.006 (.001) [.080]	.005 (.001) [.069]	.011 (.001) [.106]	.005 (.001) [.071]
Correlation in Non-persistent Innovations		.672 (.075)		.522 (.052)
Test of Fixed-Effect Model (p-value)		.000		.000
Over-identification Test (p-value)		.175		.009
Number of schools		894		894

Note: OMD estimates of parameters to describe time series in school effects $E(\delta_j' \delta_j)$. Scores are scaled so that a 1 unit change is equal to the unadjusted standard deviation in each score.

Table 4.
Decomposing the Variation in 1994 Math and Reading Test Scores

	Adjusted 5th Grade Levels		Adjusted 5th Grade Gains	
	Math	Reading	Math	Reading
Sample Size:	<i>Proportion Due to Persistent Differences Between Schools</i>			
25	0.662	0.488	0.517	0.263
50	0.760	0.623	0.584	0.355
100	0.829	0.724	0.625	0.429
Sample Size	<i>Proportion Due to Non-Persistent Differences Between Schools (Excluding Sampling Variation)</i>			
25	0.070	0.077	0.253	0.223
50	0.081	0.099	0.286	0.300
100	0.087	0.115	0.306	0.363
Sample Size	<i>Proportion Due to Sampling Variation</i>			
25	0.268	0.435	0.229	0.513
50	0.155	0.278	0.130	0.345
100	0.084	0.161	0.069	0.209

Table 5.
Performance of Schools in 1999 Identified as Being in the “Top 10%” in 1997
Based on Actual and Filtered Test Scores

5th Grade Math Gains

		<u>Based on actual 1997 Score</u>			Difference between Top 10% and the rest	Expected difference
		School not in Top 10%	School is in Top 10%	Row Total		
<u>Based on filtered prediction of 1999 Score (from 1997)</u>	School not in Top 10%	-0.016 (0.007) [N=779]	-0.022 (0.066) [N=25]	-0.016 (0.007) [N=804]	-0.006 (0.043)	0.385 (0.034)
	School is in Top 10%	0.124 (0.050) [N=25]	0.151 (0.026) [N=65]	0.144 (0.023) [N=90]	0.027 (0.052)	0.236 (0.036)
	Column Total	-0.012 (0.007) [N=804]	0.103 (0.027) [N=90]	0 (0) [N=894]	0.115 (0.024)	0.453 (0.019)
	Difference between top 10% and the rest	0.140 (0.042)	0.173 (0.059)	0.160 (0.023)		
	Expected difference	0.147 (0.013)	0.095 (0.012)	0.180 (0.007)		

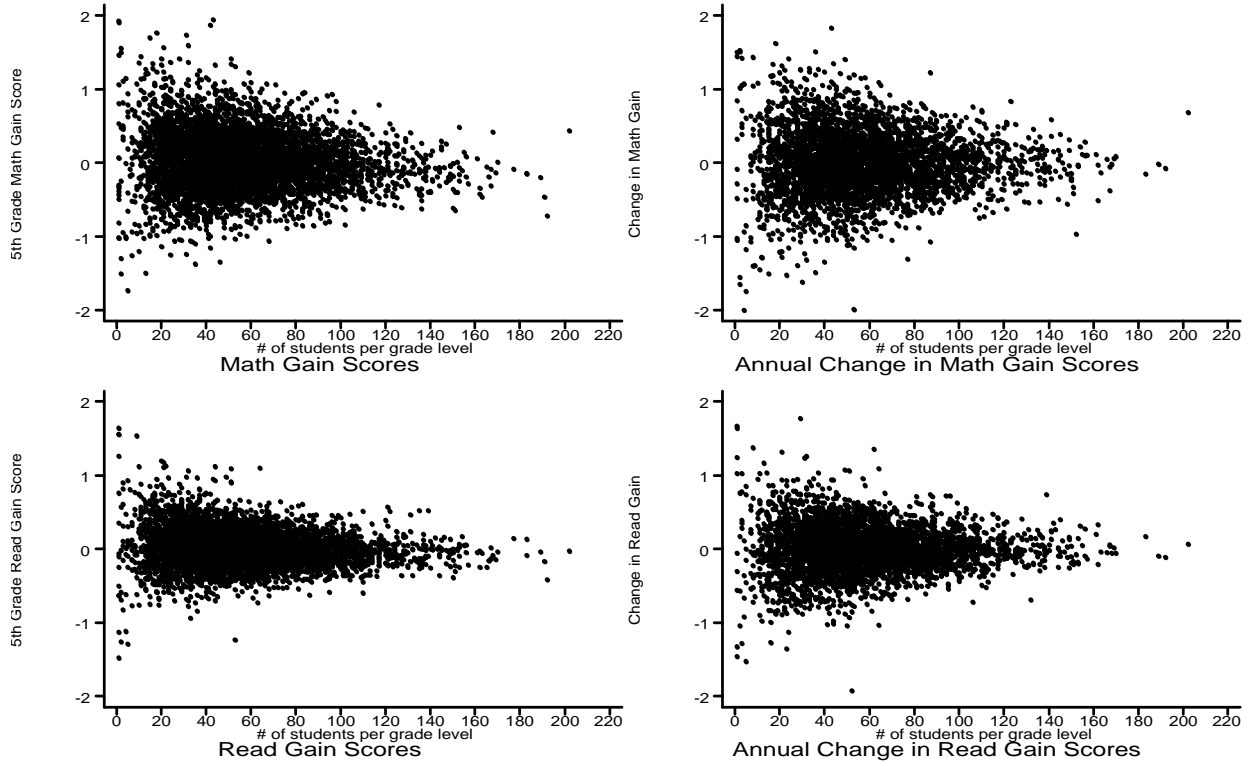
Notes: Within the box, the entries report the mean of the 5th grade math gain score in 1999, along with standard errors of these estimates and the sample size in each cell. The columns of the table use actual scores in 1997 to assign schools to “top 10%” and to calculate the expected difference between the top 10% and the rest. The rows of the table use filtered predictions of 1999 scores, based only on data from 1994-1997, to assign schools to “top 10%”.

Table 6.
Comparing the Accuracy of Alternative Forecasts
of 1998 and 1999 Test Scores

Test Score Being Predicted:	Unweighted R ² when Forecasting 1998 and 1999 Scores under Alternative Uses of 1993-97 Data					
	Predicting Scores in 1998 (1-year ahead forecast R ²)			Predicting Scores in 1999 (2-year ahead forecast R ²)		
	“Filtered” Prediction	1997 Score	Average Score 1994-97	“Filtered” Prediction	1997 Score	Average Score 1994-97
Adjusted Score						
5 th Grade Math	0.41	0.19	0.29	0.27	-0.02	0.13
5 th Grade Reading	0.39	0.13	0.33	0.31	-0.05	0.24
Gain Score						
5 th Grade Math	0.16	-0.27	0.09	0.12	-0.42	-0.01
5 th Grade Reading	0.04	-0.93	-0.12	0.04	-0.85	-0.20

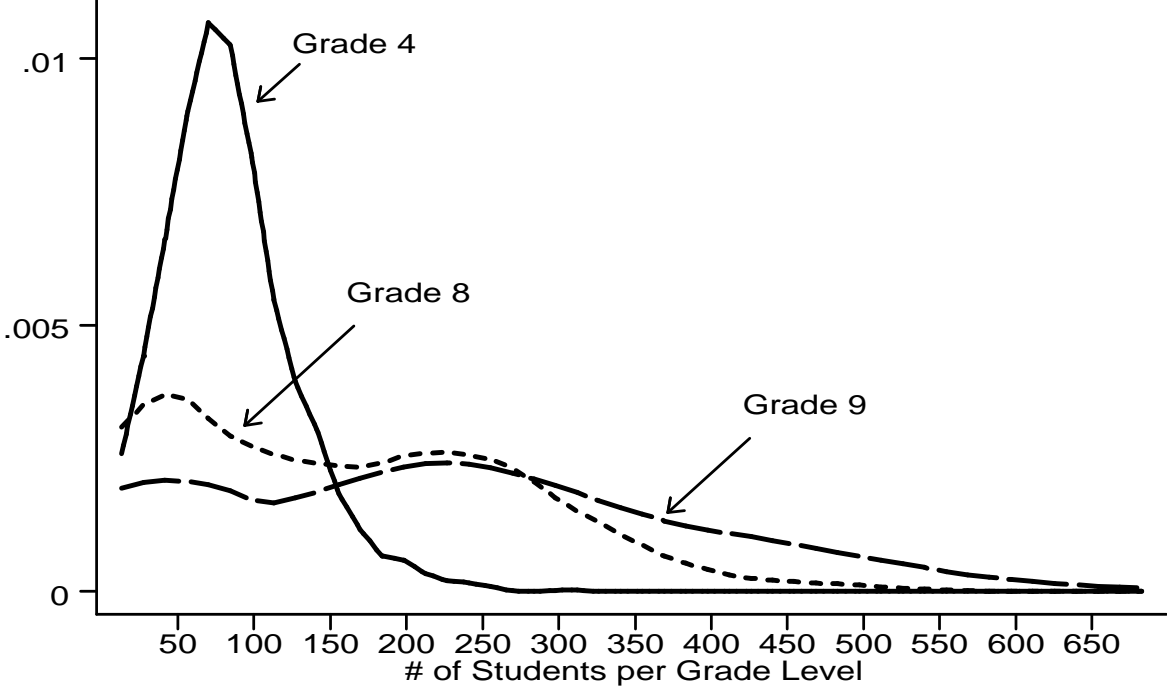
Note: The “filtered” prediction is an out-of-sample prediction, generated using only the 1993-1997 data.

Figure 1



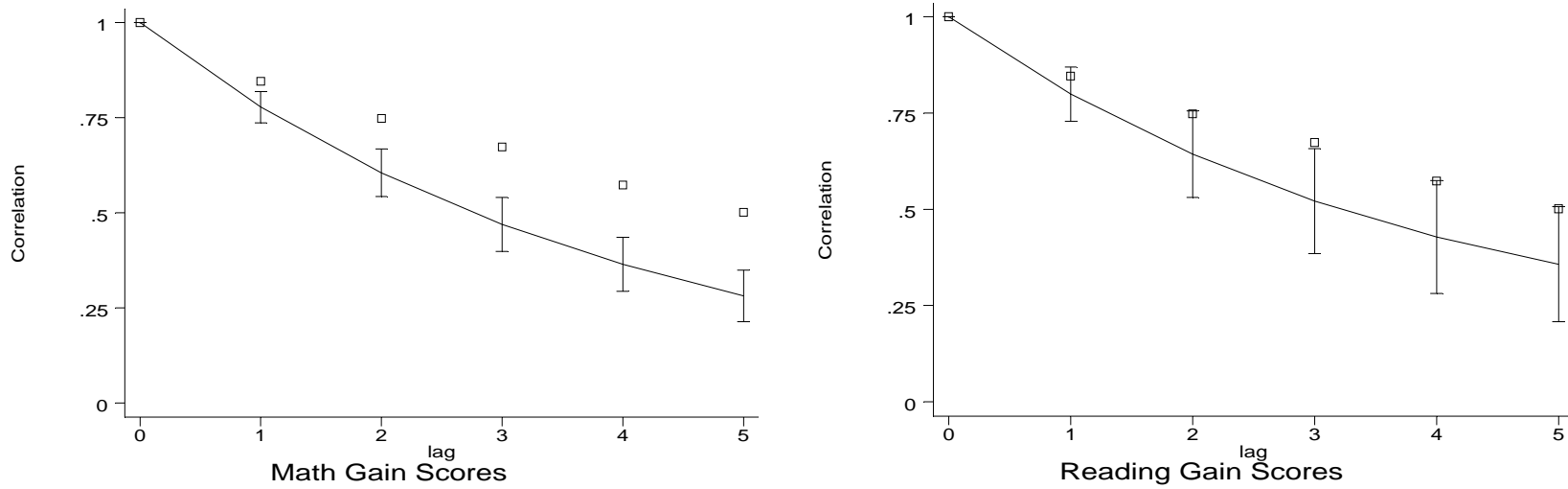
Math and Reading Gains and Changes by Sample Size

Figure 2



Distribution of School Size by Grade Level

Figure 3



Correlation in math and reading gain scores between 1999 Implied by VAR Estimates and Teacher Turnover in North Carolina (squares)