

Impediments to the estimation of teacher value added

Steven G. Rivkin
Jun Ishii

Amherst College

April, 2008

Preliminary draft

Please do not cite

The recent interest and emphasis on accountability in public schooling has placed standards and incentives at the forefront of the education policy debate. A growing numbers of districts and states are adopting pay for performance plans and accountability systems in which student test outcomes play a central role in the determination of quality ratings and in some cases compensation. Not surprisingly this has produced questions about the methods used to rate schools or teachers and the validity of rating or compensation systems based on those ratings. Although some criticisms no doubt emanate from ideological or self-interested opposition to test-based accountability, the success of these systems almost certainly depends in large part upon the acceptance of the measurement procedures as fair.

In this paper we consider potential impediments to the estimation of teacher quality caused primarily by the non-random sorting of students and teachers among communities and schools. As research discusses in great detail, the determinants of both student and teacher choices and the allocation of students among classrooms complicate efforts to isolate the contributions of teachers to learning. The discussion highlights the importance and difficulty of accounting for student differences that affect both own performance and the classroom environment and the likely advantages of an approach that accounts explicitly for prior knowledge. It also recognizes, however, that the value added framework does not address all potential impediments to consistent estimation of the quality of instruction. Specific methods mitigate some deficiencies and not others; none may resolve all potential problems. Ultimately a clearer understanding of these issues can aid in the development of effective and fair incentives for schools and teachers.

The next section describes choices by students, teachers, and principals that are primary determinants of the matching of students and teachers. The following section outlines a cumulative model of learning and considers the advantages of various estimation methods given the student-teacher matching process. Particular attention is focused on the tradeoffs between a focus on

performance differences within as opposed to between schools. In section 4, we discuss relevant findings in the literature, focusing on results from a related paper on the sensitivity of value added estimates to the way in which students are sorted into classrooms. Differences in the patterns of teacher quality estimates between schools that appear to sort students into classrooms non-randomly to those where the hypotheses of random sorting along several dimensions is not rejected at conventional significance levels provide some empirical support for the existence of problems highlighted in the conceptual discussion. Finally, the last section considers implications of our discussion for education policy and the structure of accountability and pay for performance systems.

II. Allocation of students and teachers into schools and classrooms¹

This section outlines the decision making processes of families, teachers, and principals and the potential implications of these choices for the estimation of teacher quality. Families choose community and school, possibly trading off school quality with other housing amenities. Because many data sets including administrative data have limited information on family income and other family characteristics including parental education that are likely related to both commitment to schooling and home resources available to support education, it is often difficult to control for family heterogeneity. Any failure to account adequately for family differences contaminates estimates of teacher quality unless students are randomly sorted into classrooms.

Evidence on teacher preferences suggests that teachers tend to prefer schools with higher achieving students and appear to have heterogeneous preferences regarding school location and characteristics related to student race and ethnicity.² Survey evidence suggests that principal behavior influences the probability a teacher remains in a school, likely introducing a link between teacher and

¹ This section draws extensively from Rivkin (2007).

² Boyd et al () and Hanushek, et al (2008) describe differences in teacher sorting by race.

principal quality. This likely includes the process through which principals determine classroom assignments.

A principal's objective function almost certainly influences classroom assignments and the distribution of classroom average test scores within a school. An egalitarian principal might place more disruptive children with a higher quality teacher, while a principal that desires to please the senior staff might give experienced teachers the most compliant children. These two allocation mechanisms have very different implications for the observed achievement differences among classrooms and the estimation of teacher quality.

In addition to decisions that affect the matching of students and teachers in classrooms, it is also important to recognize other parental and teacher choices that affect achievement and the estimates of teacher quality. Two in particular are the devotion of family time and resources to academic support and teach time allocation and pedagogy decisions. With regards to the former, Todd and Wolpin (2001) point out the fact that estimates of school and teacher effects incorporate family responses that vary systematically with input quality. If families tend to contribute more time and money to academic support during periods in which they perceive the teacher as inadequate or of low quality, this will tend to bias downward estimates of the variation in teacher quality and bias estimates of the quality of specific teachers toward the mean. Consequently it is not adequate to represent families as fixed inputs into the education production process, meaning that even the inclusion of student fixed effects would fail to account for potentially important family influences.

The issues of teacher time allocation and pedagogy relate much more to the interpretation of teacher fixed effects and the desirability of particular policies. Because it is learning and not simply higher test scores that matters and because mathematics and reading comprehension are not the only valued subjects, ranking on the basis of test scores alone even in the absence of specification or measurement error would not necessarily provide an ordering that would correspond to a ranking on

the basis of amount actually learned in school. First, it may take some teachers far more class time than others to produce the same achievement gains, and the opportunity cost in terms of foregone learning in other areas varies positively with the amount of class time needed per unit of achievement gain. Second, findings by Koretz (1991) suggests gearing pedagogy geared toward improvement on a specific assessment may result in less meaningful and lasting knowledge acquisition.

All in all, the purposeful nature of these choice processes almost certainly introduces correlations among teacher quality, school quality, and family and student characteristics which complicate efforts to identify teacher effects on achievement. The following section develops an empirical model of achievement based on the notion that learning is a cumulative process. The discussion highlights the extent to which specific methods address complications introduced by the choices of students, teachers, and principals as well as test measurement error.

III. Cumulative Model of Learning

Equation (1) models achievement of student i in grade G in classroom c in school s in year y as a function of student skill (α_{iGy}), family background (X), peer composition in classroom c during year y (P), school factors specific to grade G in year y – including resources, principal quality, and school or district determined curriculum – (S), teacher quality (T), and a random error (e). Without loss of generality, think of each of these terms as scalar indexes of the respective characteristics that increase in value as the characteristic becomes more conducive to achievement. For example, a higher value of P indicates a better peer composition (perhaps fewer disruptive students). Therefore all of the parameters are non-negative, as higher skill, family characteristics that support achievement, better peer compositions, better schools, and higher teacher quality all raise test scores.

$$(1) A_{iGcsy} = \alpha_{iGy} + \beta X_{iGy} + \tau P_{Gcsy} + \delta S_{Gcsy} + \lambda T_{Gcsy} + e_{iGcsy}$$

Importantly, teacher quality is represented in the above regression equation by a full set of indicators for each teacher-year combination, permitting teacher effectiveness to vary with experience and other factors that change over time. This implies that teacher quality is not directly observed but rather estimated by netting out the contributions of student skill (α_{iGy}), family background (X), peer composition (P), and school factor (S) from the average achievement of students taught by the particular teacher (or by the particular teacher for that year, in the case of teacher by year indicators).

In the absence of random assignment, unobserved peer and school factors for a given class (G_{csy}) will likely confound estimates of the quality of the teacher assigned to that class; some of the achievement for students in that class will be attributed to the teacher rather than the unobserved peer and school characteristics of that class. Principal assignment of better teachers to classrooms with better students (or worse students, if seeking to equalize achievement across classes), better teachers gravitating toward higher resource schools, families with the most educational concerns and most resources to support children educationally moving to the school districts with the best teachers all complicate the estimation of teacher value added to achievement, as teacher quality becomes bundled with characteristics of students or schools. It is clear that families do not choose communities and schools at random, making comparisons among schools quite difficult. Moreover, it is quite likely that principals and others involved in the assignment of students to classrooms consider numerous factors.

Of course if teachers were randomly assigned to different students and classes, many times, the average achievement of students taught by the teacher could be used to rank teachers. The random multiple assignments would ensure that the contribution of non-teacher factors to average student achievement is the same, in expectation, across teachers. So differences in the average student achievement across teachers could be attributed to differences in teacher quality. The precision of this attribution increases with the number of random student/school assignments made to each teacher, as

more (random) assignments better ensure the actual comparability of the non-teacher factors to student achievement across teachers. No sophisticated regression analysis is necessary as the simple act of averaging largely controls for the contribution of other factors.

When non-random assignments introduce correlation between teacher quality and other factors, averaging is insufficient. A method, such as regression analysis, that isolates the effects of teacher quality from other influences, must be implemented. The desirability of any particular approach depends upon the extent to which it accounts for the potential confounding factors. For the linear regression approach, teacher quality is properly isolated to the extent that other confounding factors are properly included and specified as explanatory variables in the regression. Either omission or misspecification of important confounding factors corrupts the estimate of the teacher quality obtained from the regression.

An important potential source of such corruption is student heterogeneity. Student skill, like teacher quality, is not directly observed and must be inferred using some model of student ability. Because student differences in cognitive skills evolve over time with educational experiences at home, in school, and in the community, student skill needs to be modeled using the full history of family, teacher, peer, and community influences on the student. One possible way to model this cumulative learning is to use the following specification for student skill

$$(2) \alpha_{iGy} = \beta \sum_{g=1}^{G-1} \theta^{G-g} X_{igy-g} + \delta \sum_{g=1}^{G-1} \theta^{G-g} P_{gcsy-g} + \tau \sum_{g=1}^{G-1} \theta^{G-g} S_{gcsy-g} + \lambda \sum_{g=1}^{G-1} \theta^{G-g} T_{gcsy-g} + (\gamma + \sum_{g=1}^{G-1} \theta^{G-g} \gamma_i)$$

A good teacher likely raises achievement in the current year and subsequent years by increasing the stock of knowledge, and a very supportive parent does the same. Notice that factor effects (and knowledge) are assumed to depreciate at a geometric rate, meaning that a teacher or peer's effect on

test scores diminishes with time such that a good 4th grade teacher has a larger effect on 4th grade score than on 5th grade score. The equation does not specify the rate of depreciation.³ If $\theta=1$ the effects of prior experiences persist fully into the future, while if $\theta=0$ prior experiences have no effect on current achievement. It is highly likely that the actual knowledge depreciation rate lies between 0 and 1.

A value added regression of achievement in grade G on achievement in grade G-1, family, school, and peer characteristics, and a full set of indicators for each teacher provides a natural way to account for prior influences and estimate teacher effects on achievement (the dummy variable coefficients).⁴ Rewriting equations (1) and (2) for grade G-1 and year y-1 illustrates how the inclusion of $A_{iG-1,csy-1}$ as an explanatory variable with parameter θ in a regression with achievement in grade G as dependent variable potentially controls for the full set of historical factors.

$$(3) A_{iGcsy} = \theta A_{iG-1csy-1} + \gamma_i + \beta X_{iGy} + \tau P_{Gcsy} + \delta S_{Gcsy} + \lambda T_{Gcsy} + e_{iGcsy}$$

In the absence of test measurement error ($e_{iG-1csy-1} = 0$) only the contemporaneous ability effect γ_i remains unaccounted for regardless of the rate knowledge depreciates (assuming the source of knowledge does not affect the rate of depreciation). And since this fixed ability component is likely to be highly correlated with lagged achievement, controlling for lagged achievement differences almost certainly remove much of the variation in contemporaneous ability as well. Consequently in the absence of measurement error the value added specification would appear to provide an excellent method for capturing skill differences that contribute to variation in achievement among classrooms.

³ An alternative value added specification is to use the difference in scores between grades G and G-1 as the dependent variable thus imposing the assumption of $\theta=1$. As Rivkin (2006) demonstrates, this more restrictive framework will tend to bias downward differences among teachers in the absence of student fixed effects and bias upward differences among teachers if student fixed effects are included.

⁴ See Hanushek (1986) for a discussion of value added models.

IV. Estimation of Teacher Fixed Effects

Lagged achievement can help eliminate, or at least mitigate, omitted variable and misspecification concerns associated with student skill. But the estimation of teacher quality can still be confounded by omitted variables and/or misspecification of the other factors – family circumstances (X), peer composition (P), school characteristics (S) as well as test measurement error. These omitted or misspecified confounding factors lead to unobserved heterogeneity, across students, in non-teacher contribution to achievement which, in turn, make it difficult to isolate the contribution of teacher quality. As discussed above, this unobserved non-teacher differences may only be across schools or also within schools. In the case of the former, teacher quality may still be estimated consistently relative to other teachers in the same school or even same school, grade, and year. Notable sources of “across school” unobserved heterogeneity include the quality of the principal, the extent to which the curriculum for grade G lines up with the state test, and the level of student disruption. But in the latter case – unobserved heterogeneity within school – a “within school” estimator will suffer from omitted variables bias. This correlation would exist if students were purposefully sorted into classrooms, parents react to teacher quality in choosing their level of effort, or if school specialists such as a reading or math teacher provided compensatory teaching.

Of particular concern are peer and school factors that vary only at the classroom, grade, or school level. With teachers observed teaching only one class each year, the inclusion of teacher by year fixed effects necessarily precludes, from the regression, any other explanatory variables that do not vary within classroom. The influence of the precluded peer and school factors become bundled with that of teacher quality in the estimated teacher fixed effect. It is possible to include a set of observable peer and school characteristics in a second stage regression of teacher fixed effects on these variables, but it is unlikely that the limited observed characteristics will capture fully relevant differences in school leadership, students’ behavior, and other determinants of achievement.

IVa. Decomposition of the teacher fixed effect

Both bias and sampling variability can cause the teacher fixed effect estimates to deviate from the actual teacher contributions to learning, a point highlighted by Kane and Staiger (2001). Equation (4) presents the teacher j fixed effect as a sum of the true teacher effect, the average student contribution to her own learning (subscript i), from her peer environment (subscript p), from the contribution of her school factors (subscript s), and random sampling error:

$$(5) \quad \hat{t}_j = t_j + \hat{\varepsilon}_i + \hat{\varepsilon}_p + \hat{\varepsilon}_s + \hat{\varepsilon}_n$$

Clearly the estimate deviates from the true teacher effect, but the appropriate remedy depends upon the relative importance of bias and sampling variability. Consider first the case of no bias. In this situation shrinkage estimation or the correlation across multiple teacher quality estimates for the same teacher can be used to produce a consistent estimate of the true variance in teacher quality.⁵ Alternatively, the correlation among multiple fixed effect estimates for the same teacher can be used. Maintaining the assumption of random error components, the variance of \hat{t}_j is the sum of the true variance of quality (t_j) and the variances of the errors.

Moreover, the expected value of the correlation of fixed effect estimates for the same teachers across years, $E(r_{12})$, is:

$$(6) \quad E(r_{12}) = \text{var}(t) / \text{var}(\hat{t})$$

Multiplication of the estimated variance of \hat{t} by the year-to-year correlation thus provides an estimate of the overall variance in teacher quality corrected for the random contributions of test error and the other factors. Importantly, this approach addresses problems related to both the noisiness of tests as measures of learning and any single year shocks (either purposeful or random) in classroom average student quality.

⁵ Jacob and Lufgren (), Kane and Staiger (2001), Aronsen et al (2003), Rockoff (2007), and others use estimates of the error variance to adjust raw fixed effect estimates.

In this case the removal of all non-persistent variation also eliminates some portion of the true variation in quality resulting from changes in actual effectiveness. Moreover, the estimate of quality based on within school variation may also change over time even in the absence of any true difference in performance, because the within school estimates of quality are defined relative to other teachers. Any turnover can alter a teacher's place in the quality distribution in her school. Therefore, by considering just the persistent quality differences, some true systematic differences in teachers are masked by a varying comparison group and are treated as random noise.

With bias and thus persistent correlation between confounding factors and teacher quality efforts, error adjusted estimates will not produce meaningful measures of the contributions of teachers. The potential problem of confounding differences across schools which make it difficult to separate teacher and school effects has led to a focus on within school differences. But recent work by Rothstein (2007) and Clotfelter et al (2006) provide evidence of substantial within school sorting. This raises questions about the information contained in within school as well as between school estimates of teacher quality, an issue to which we turn in the empirical section.

IVb. Model of Test Score Gains

Because of its prevalence in research and school accountability and potential value in understanding the role of non-random teacher-student matching, we now briefly discuss some implications of using a test score gain rather than a lagged achievement model to account for differences in past determinants of current achievement. In the test score gain model, instead of including lagged achievement as a regressor, the test score gain (current score minus prior year score) is used as a dependent variable:

$$(7) A_{iGcsy} - A_{iG-1csy-1} = \gamma_i + \beta X_{iGy} + \tau P_{Gcsy} + \delta S_{Gcsy} + \lambda T_{Gcsy} + \eta_{iGcsy}$$

An important advantage of this specification is that it eliminates potential biases in the estimates of teacher fixed effects and other coefficients resulting from measurement error in the lagged achievement measure. The noisier the measure of lagged achievement the less it captures variation in student academic preparation. Although additional lagged test scores or the instrumenting of lagged score with twice lagged score or another score can also mitigate such biases, these measures are not feasible in many instances.

Despite this potential advantage, the gains model described in equation (7) likely introduces specification error into estimation of school and teacher effects by implicitly imposing the assumption of no knowledge depreciation ($\theta=1$). Consider a regression model in which we assume no factors other than past value of teacher quality are related to both current teacher quality and achievement, the variance of teacher quality is constant across grades, and the correlation (ρ) in teacher quality across grades is constant regardless of the number of grades apart.⁶ Equation (5) presents achievement gain in grade G as a function of teacher quality (TQ) in grade G and a regression error.

$$(8) \quad A_{iG} - A_{iG-1} = TQ_G \beta_{gain} + gain\ error =$$

$$TQ_G \beta_{gain} + \beta \sum_{g=1}^G (\theta^g - \theta^{g-1}) TQ_{G-g} + error$$

Notice that the error includes lagged values of teacher quality but not other variables in order to highlight the implications of persistent sorting into schools and classrooms.

The expected value of the estimate of β_{gain} equals⁷

$$(9) \quad E(\hat{\beta}_{gain}) = \beta - \beta \left(\frac{(1 - \theta^G) \rho}{\text{var}(TQ)} \right)$$

⁶ Rivkin (2006) discusses specification error in value added models in detail.

⁷ Given the assumption that $\text{cov}(SC_i, SC_j)$ is constant regardless of the number of grades apart, all terms cancel except for one grade G-1 term and one grade 0 term.

The magnitude of any bias depends on both θ and ρ . Not surprisingly given the structure of the model, bias decreases as the θ increases and disappears if there is no loss of knowledge from year to year, i.e. $\theta=1$. As is the case with a model that does not control for lagged achievement, random teacher assignment or other methods that isolate the component of the school characteristic in grade G that is uncorrelated with the value of the characteristic in other grades produce unbiased estimates of β regardless of the value of θ . Importantly, the inclusion of a student fixed effect into a gains model does not eliminate the bias as long as there is correlation among teacher quality across grades.

More generally, when there is some knowledge depreciation ($\theta < 1$), the regression error, η_{iGcsy} , in equation (7) is implicitly the sum of e_{iGcsy} and $(\theta-1)A_{iG-1csy-1}$. This can be seen by comparing equations (3) and (4). If the family, peer, school, or teacher effects for a student this year are correlated with those for that student last year, then the contemporary value of those effects are also correlated with the contemporary regression error due to $A_{iG-1csy-1}$. This results in the standard endogeneity bias.

V. Understanding Estimates of Teacher Quality

There has been extensive academic and policy work that estimates and makes use of information on teacher fixed effects despite the existence of ongoing concerns about the difficulties of identifying precisely the contributions of teachers to achievement. On the one hand, recent work by Rothstein (2007) suggests that the potential problems are quite severe and raise questions about the use of such estimates in merit compensation and school accountability programs. On the other hand, a comparison of estimates from studies using very different approaches suggests that difficulties in separating teacher contributions from those of confounding factors are clearly a problem but that there is not strong evidence that factors other than teachers account for most of the observed variation

among teachers or that differences in teacher quality are not important.

Most of the estimates of the variance in teacher quality come from analyses that estimate teacher fixed effects directly from data that matches students with their teachers. But Rivkin et al (2005) use grade level data and the link between within school achievement variation across cohorts and teacher turnover to back out an estimate of the variation in teacher quality. Because purposeful sorting should have little or no effect on cohort differences in achievement gains, the finding of significant differences in teacher quality cannot be explained by non-random classroom sorting. However, the fact that the magnitude of the estimate in standard deviation units falls substantially below most estimates generated by studies based on direct estimation of teacher fixed effects is consistent with the hypothesis that purposeful sorting inflates estimates of the variance of teacher quality even in studies that focus on differences within schools.

In an effort to learn more about this issue we have investigated patterns of teacher fixed effect estimates for a large Texas public school district. The analysis and findings are reported in Hanushek et al (2008). We discuss some of the results here to illustrate our discussion above.

Va. Sorting and estimates of the variance in teacher quality

Hanushek et al (2008) estimates systematic differences in value added by teacher transition status and the sensitivity of these estimates and estimates of the variance of teacher quality to the ways in which students and teachers are matched and to the structure of the value added model. Prior to reporting the results we describe the data and specifications.

The Texas Schools Project Micro-Data contains administrative records on students and teachers collected by the Texas Education Agency (TEA) from the 1989-90 school year through 2001-2002. The data permit the linkage of students over time and of students and teachers in the same school,

grade, and year. The statewide data do not match students and classroom teachers, but those matches have been provided for the single large district.

The student background data contain a number of student, family, and program characteristics including race, ethnicity, gender, and eligibility for a free or reduced price lunch (the measure of economic disadvantage), classification as special needs, and classification as limited English proficient. Students are annually tested in a number of subjects using the Texas Assessment of Academic Skills (TAAS), which was administered each spring to eligible students enrolled in grades three through eight. These criterion referenced tests evaluate student mastery of grade-specific subject matter, and this paper presents results for mathematics. Test scores are converted to z scores using the mean and standard deviation for the entire state separately for each grade and year to account for the effects of test score inflation and other changes to the tests.

In this paper we study students and teachers in grades 4 through 8 for the school years 1995/1996 to 2000/2001. We eliminate any student without valid test scores or other missing data and classrooms with fewer than five students with non-missing data. Separate teacher fixed effects for each year that a teacher appears in the data are estimated from a specification that controls for the aforementioned student characteristics and standardized math score in the prior grade.

In order to analyze the various impediments to estimating teacher value-added, the estimated teacher year fixed effects and corresponding measures of teacher quality variance are compared across different model specifications and samples. We explore possible misspecification bias stemming from imposing no knowledge depreciation by producing full set of estimates for both the lagged achievement and test score gain specifications of student skill. To investigate the influence of unobserved family, peer, school effects, we calculate both the overall variation in the district and the variation within schools, grades, and years (obtained by demeaning by the average fixed effect for the school-grade-year). Differences between the overall and within variation suggest the presence of

important unobserved heterogeneity within school, grade, year.

We address the issue of purposeful sorting by dividing the sample into observations from schools/grade/year that exhibit significant signs of sorting and from those that do not. Full set of estimates are obtained for each sub-sample. Differences in the teacher value-added estimates and variance suggest the influence of sorting. Two methods are used to divide the sample. The first method follows in the spirit of Clotfelter et al (2006) and divides the sample based on whether observations from the school/grade/year reject the null hypothesis of no difference in the mean pretest score across classrooms. The second examines the transitions of students who remain in the school in grades $g-1$ and g and tests for the independence of the classroom allocation in the two grades using a chi square test. For each method, the observations from school/grade/year that reject the null hypothesis are considered observations affected by purposeful sorting (“non-random”).

Lastly, we consider a falsification exercise similar to Rothstein (2007). Estimates for each of the model specification and samples are obtained where, indicators for current year teacher are replaced by indicators for subsequent year teacher for each student i . Because the subsequent year teacher cannot directly affect current achievement, a finding of significant differences by subsequent year teacher provides evidence that persistent differences in schools or peers confound estimates of teacher quality. With respect to the “random” and “non-random” sorting sample division, students are divided on the basis of whether the hypothesis of non-random sorting in grade $g+1$ (rather than g) is rejected.

Vb. Preliminary Results

Our current estimates find evidence suggestive of non-random sorting confounding teacher quality estimates. Estimates of the variance in teacher quality tend to be uniformly larger for schools in which the hypothesis of random sorting is rejected. This is driven by much larger correlations across

years in the estimates for the same teacher in these schools. In terms of within school variances, the difference exceeds 60 percent (0.021 to 0.013), when the mean pretest score is used to divide the sample, and approaches 40 percent (0.015 to 0.011), when the student transitions are used to divide the sample. The larger variance estimates in the pretest sample may result from the larger sample sizes that reduce the error variances.

We also find within school variance estimates from the test score gain model to be much smaller than those from the lagged achievement model for the “non-random” sub-sample but not for the “random” sub-sample. In the case of the “non-random” sub-samples, the variance falls from 0.021 to 0.017 with the pretest division and from 0.015 to 0.009 with the division by student transitions. In the case of the “random sample”, the variance estimate actually increases from 0.13 to 0.18 in the pretest samples and remains constant at 0.11 in the transition samples. Specification error, due to the incorrect assertion of no knowledge depreciation, emerges only if current teacher quality is related to prior determinants of achievement. Therefore, one would expect the error to vary with the degree of correlation. Our finding of no change in the estimates for the “random” sample but a substantial decrease in the “non-random” sample further supports the notion that confounding factors are present in that latter group of schools. Our estimates of the overall variance all decline when the test score gain model is used, consistent with the existence of substantial sorting across schools.

Our falsification exercise using the “subsequent year teacher” counterfactual further corroborates the confounding influence of purposeful sorting. The counterfactual within school variance estimates for future teachers using the lagged achievement model are roughly 60 percent smaller than the variance estimates using the actual teachers for grades in the non-random allocation category (0.018 to 0.008 and 0.022 to 0.008 in the pretest and transition samples respectively). In contrast, the estimates of the counter-factual within school variances approach zero for the sample of schools in the “random” sample, while the variance estimates based on actual teachers equal 0.017 and

0.016). This is precisely what would be expected if there were in fact significant differences in teacher quality and students were randomly distributed among classrooms. Random sampling variability should not lead to persistent errors, while true teacher effectiveness should exhibit a high degree of serial correlation from year to year.

Not surprisingly, the difference between estimates of the counterfactual and actual variances in teacher quality are much smaller in the case of the overall variance, as persistent school factors inflate estimates for both current and future teachers. In the case of the non-randomly sorted sample the counterfactual variance estimates are roughly 80 percent as large, while for the “random” sample the estimates range from roughly 80 percent to just under 10 percent as large.

All in all, these findings support both the belief in substantial variation in teacher quality and the hypothesis that unobserved student or peer factors contaminate estimates of teacher quality in schools that do not allocate students randomly among classrooms despite the use of value added specifications. Because sorting is much more prevalent in middle school than in the early grades, the process of ranking by value added is less likely to be fair in middle school than in elementary school.

VI. Value added and policy

The myriad factors that influence cognitive growth over extended periods of time, the purposeful sorting of families and teachers into schools and classrooms, compensatory behavior on the part of families, and the imperfections of tests as measures of knowledge complicate efforts to estimate the variance of teacher effectiveness and rank teachers according to quality of instruction. Even within school rankings are subject to biases and the vagaries of sampling variability. Along with possible distortions of classroom time allocation and teaching methods in an effort to increase scores, these problems raise concerns about the use of tests for high stakes purposes.

Yet the elimination of explicit learning incentives entirely seems even more problematic, as the

tests do provide goals for students and teachers and information about teacher performance. Moreover, in the hands of a principal armed with contextual knowledge about classroom environment, test results also provide important information about teacher effectiveness.

If test scores are to be used to rank schools and teachers, school average value added overall or in a single graded would appear to constitute a more fruitful approach to provide incentives to teachers and school administrators and information to families despite any difficulties in fully accounting for potential confounding factors at the family or community level. Although cooperation among teachers is a worthy consideration, we believe there are at least two more important reasons to focus on school level outcomes. First, this does not impede principals from considering the strengths and weaknesses of teachers and students in the classroom allocation process. Second, this provides a strong set of incentives for school leaders who make the key personnel, spending, and curricular decisions and should be held accountable for their actions.

Consideration of potential limitations can also highlight the tradeoffs of specific approaches and facilitate prudent use of tests in the mentoring, evaluation, and compensation of teachers and appropriate steps to maximize their value. For example, although averaging teacher performance over a number of years can reduce the influence of measurement error and make the teacher effect estimates more precise, it also averages over real year-to-year differences in effectiveness resulting from improvement with experience, personal circumstances, experimentation with new pedagogies, etc. Thus there is a tradeoff between obtaining more precise estimates and recognizing that teacher quality is not a fixed characteristic. From the point of view of a teacher, averaging dampens the incentives to improve since a bad year will weigh down performance in the future and lessen the reward from turning things around.

The development of superior tests that span the entire range of school learning objectives would also be a worthy goal. Although it would violate the annual testing requirements of NCLB, the

adoption of end of comprehensive end of school examinations similar to those used in many countries in Europe and Asia could provide high, clear targets for students and schools. There would be less information, particularly for teachers in grades far from the tested grade. However, such a high stakes test would place plenty of pressure on schools to raise quality throughout the grade distribution and provide incentives for administrators to behave in ways conducive to high quality.

Bibliography

- Aaronson, Daniel, Lisa Barrow, and William Sander. 2003. "Teachers and Student Achievement in the Chicago Public High Schools." WP 2002-28, Federal Reserve Bank of Chicago (June)
- Ballou, D., W. Sanders and P. Wright, "Controlling for Student Background in Value-Added Assessment of Teachers," *Journal of Educational and Behavioral Statistics*, 29, no. 1 (spring, 2004)
- Clotfelter, C.T., H.F. Ladd and J.L. Vigdor. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources* 41.4 (Fall, 2006): 778-820.
- Hanushek, Eric, "The economics of schooling: Production and efficiency in public schools." *Journal of Economic Literature* 24, no.3 (September 1986)
- Hanushek, E., J. Ishii, and S. Rivkin, "Estimating Teacher Quality with Purposeful Parents, Principals, and Teachers." Unpublished manuscript (April 2008)
- Hanushek, Eric A., John F. Kain, and Steve G. Rivkin. 2004. "Why public schools lose teachers." *Journal of Human Resources* 39,no.2:326-354.
- Kane, Thomas J., and Douglas O. Staiger, "Improving school accountability measures." WP 8156, National Bureau of Economic Research (2001)
- Koretz, D. and others, "The Effects of High-Stakes Testing on Achievement: Preliminary Findings about Generalization across Tests," paper presented annual AERA meeting (1991)
- McCaffrey, D., J. R. Lockwood, D. Koretz, T. Louis, and L. Hamilton, "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics*, 29, no. 1 (spring, 2004)
- Rivkin, S., "Cumulative Nature of Learning and Specification Bias in Education Research," unpublished manuscript (January, 2006)
- Rivkin, S., "Value Added Analysis and Education Policy, Calder Brief (September, 2007)
- Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, schools, and academic achievement." *Econometrica* 73,no.2 (March).
- Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94,no.2 (May):247-252.
- Rothstein, J., "On the Identification of Teacher Effects on Student Achievement: Do Value Added Models Add Value?, unpublished manuscript (September 2007)
- Todd, P. and K. Wolpin, "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal* (February, 2003)