

Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates

by

Thomas J. Kane
Harvard Grad School of Education

Douglas O. Staiger
Dartmouth College

PRELIMINARY DRAFT

Please do not cite or circulate without permission.

March 17, 2008

This analysis was supported by the Spencer Foundation. Initial data collection was supported by a grant from the National Board on Professional Teaching Standards to the Urban Education Partnership in Los Angeles. Steve Cantrell and Jon Fullerton collaborated in the design and implementation of an evaluation of the National Board for Professional Teaching Standards applicants in LA. The authors wish to thank a number of current and former employees of LAUSD, including Ted Bartell, Jeff White, Glenn Daley, Jonathan Stern and Jessica Norman. From the Urban Education Partnership, Susan Way Smith helped initiate the project and Erin McGoldrick oversaw the first year of implementation. An external advisory board composed of Eric Hanushek, Daniel Goldhaber and Dale Ballou provided guidance on initial study design. Jeffrey Geppert helped with the early data assembly. Eric Taylor provided excellent research support for the analysis in this paper.

Abstract

For more than three decades, analyses of non-experimental data have reported considerable heterogeneity in teacher impacts on student achievement. We use data from a random-assignment experiment in Los Angeles Unified School District to test the validity of various non-experimental specifications. To do so, we first generated estimates for a group of teachers during a pre-experimental period and then tested the validity of those teacher-level estimates in predicting student achievement following random assignment. First, although the validation exercise revealed a slight advantage for a specification including prior student test scores and mean classroom characteristics as covariates (both in terms of bias and mean-squared error), several alternative non-experimental specifications were able to predict experimental outcomes. Second, although there is greater signal variance in teacher effects on math performance than on English language arts, the impacts on math performance seem to fade-out more rapidly. Specifically, half of the impact on math performance faded after one year and another half disappeared the following year; the impacts on English language arts are smaller in magnitude but more long-lived. Third, the same pattern of substantial fade-out in math (and less fade-out in English language arts) is apparent in the non-experimental data.

Introduction

For more than three decades, research in a variety of school districts and states has suggested considerable heterogeneity in teacher impacts on student achievement. The finding appears to corroborate the popular perception that one or two teachers can have a huge impact on one's life.¹ However, perhaps *because* such results are so consistent with conventional wisdom, their empirical grounding is rarely questioned. Unfortunately, the finding rests almost entirely on non-experimental variation in student achievement. However, as several recent papers remind us (Andrabi, Das, Khwaja and Zajonc (2008), McCaffrey et. al. (2004), Raudenbush (2004), Rothstein (2007), Rubin, Stuart and Zannutto (2004), Todd and Wolpin (2003)), the statistical assumptions required for the identification of causal teacher effects with observational data are extraordinarily strong-- and rarely tested. Over time, teachers may be consistently assigned classrooms of students that differ in unmeasured ways—such as consisting of more motivated students, or students with stronger unmeasured prior achievement or more engaged parents—that result in varying student achievement gains. If so, rather than reflecting the talents and skills of individual teachers, estimates of teacher effects may reflect principals' preferential treatment of their favorite colleagues, or ability-tracking based on information not captured by prior test scores, or the advocacy of engaged parents for specific teachers.

In this paper, we rely on an experiment in which 78 pairs of classrooms (156 classrooms and 3194 students) were randomly assigned between teachers in the school years 2003-04 and 2004-05.² We generate traditional “value-added” estimates for these teachers using data from a pre-experimental period, and employing a variety of non-experimental

¹ Autobiographies of successful people frequently cite the impact of an individual teacher. We presume that teachers also play a role in personal failure, but those stories are less often told.

² We began with 92 pairs of teachers and 3799 students. However, we lost 14 pairs when prior-value added estimates were missing for at least one of the teachers in the pair.

specifications. For the experiment, principals in each of the schools were asked to draw up two classrooms they would be equally happy to have assigned to each of the teachers in the pair.³ The school district office then randomly assigned the classrooms to the teachers. We then test each specification's ability to predict within-pair student achievement differences one, two and three years later.

We report the following results. First, although the experimental validation revealed a slight advantage for a specification including prior test scores and mean peer characteristics as covariates (both in terms of bias and mean-squared error), several alternative non-experimental specifications were able to predict differences in mean student outcomes within each pair. For instance, although our results suggest that raw mean end-of-year test scores (unadjusted for student covariates) overstate teacher differences, and that differencing out student fixed effects in test score levels understates teacher differences, one could not reject the hypothesis that both of these crude measures of a teacher's impact were able to predict differences in student achievement during the experiment..

Second, although there seems to be greater signal variance in teacher effects on math performance than on English language arts, the impacts on math performance seem to fade-out more rapidly: Half of the impact of being assigned a "high value-added teacher" faded by the second year and much of the remainder disappeared the following year. Although the absolute variation in teacher effects is smaller in English language arts than in math, our results suggest that those impacts are more long-lived.

Third, a similar pattern of fade-out is observed in the non-experimental data. For those teachers who were not in the experimental schools, we estimated teacher effects

³ The sample was designed to measure differences in teacher impacts associated with scores on the National Board for Professional Teaching Standards' certification process. As a result, half of the teachers had applied for National Board certification.

during the pre-experimental period. We then tested validity of these teacher impact estimates from earlier cohorts in predicting student performance for a later cohort of students assigned to each teacher 2004-05. For those not included in the experimental sample, we studied the relationship between the 2004-05 teacher assignment and performance in the spring of 2006 and 2007. (This was analogous to our analysis of fade-out in the experimental sample.) The degree of fade-out in the non-experimental sample was consistent with what we observed in the experimental sample, although somewhat more pronounced.

Related Literature

Although there have been a number of studies using non-experimental data to estimate teacher effects (for example, Armour (1971), Hanushek (1976), McCaffrey et. al. (2004), Murnane and Phillips (1981), Rockoff (2004), Hanushek, Rivkin and Kain (2005), Jacob and Lefgren (2005), Aaronson, Barrow and Sander (2007), Kane, Rockoff and Staiger (2006), Gordon, Kane and Staiger (2006)), we were able to identify only one previous study using random assignment to estimate the variation in teacher effects. In that analysis, Nye, Konstantopoulous and Hedges (2004) re-analyzed the results of the STAR experiment in Tennessee, in which teachers were randomly assigned to classrooms of varying sizes within grades K through 3. After accounting for the effect of different classroom size groupings, their estimate of the variance in teacher effects was well within the range typically reported in the non-experimental literature. However, the study was not designed to provide a validation of non-experimental methods. The heterogeneity of the teachers in those 79 schools may have been non-representative or rivalrous behavior induced by the experiment itself-- or simple coincidence-- may have accounted for the similarity in the estimated

variance with the non-experimental literature. Because they had only the experimental estimates for each teacher, they could not test whether non-experimental techniques would have identified the *same* individual teachers as effective or ineffective. Yet virtually any use of non-experimental methods for policy purposes would require such validity.

Description of the Experiment

The experimental portion of the study took place over two school years: 2003-04 and 2004-05. The initial purpose of the experiment was to study differences in student achievement among classrooms taught by teachers certified by The National Board for Professional Teaching Standards (NBPTS)—a non-profit that certifies teachers based on a portfolio of teacher work (two 20 minute videotapes, lesson plans, examples of student work and short essay responses). Accordingly, we began with a list of all National Board applicants in the Los Angeles area (identified by zip code). LAUSD matched the list with their current employees, allowing the team to identify those teachers still employed by the District.

Once the National Board applicants were identified, the study team identified a list of comparison teachers in each school. Comparison teachers had to teach the same grade and be part of the same calendar track as the National Board Applicants. In addition, the NBPTS requires that teachers have at least three years of experience before application. Since prior research has suggested that teacher impacts on student achievement grow rapidly during the first three years of teaching, we restricted the comparison sample to those with at least three years of teaching experience.

The sample population was restricted to grades two through five, since students in these grades typically are assigned a single instructor for all subjects. Although

participation was voluntary, school principals were sent a letter from the District's Chief of Staff requesting their participation in the study. These letters were subsequently followed up with phone calls from the District's Program Evaluation and Research Branch (PERB). Once the comparison teacher was agreed upon and the principal agreed to participate, the principal was asked to choose a date upon which the random assignment of rosters to teachers would be made. (Principals either sent PERB rosters or already had them entered into LAUSD's student information system). On the chosen date, LAUSD's PERB in conjunction with the LAUSD's School Information Branch randomly chose which rosters to switch and executed the switches at the Student Information System at the central office. Principals were then informed whether or not the roster switch had occurred. Seventy eight valid pairs of teachers, each with prior non-experimental value-added estimates, were included in the present analysis.

Once the roster switches had occurred, no further contact was made with the school. Some students presumably later switched between classes. However, 85 percent of students remained with the assigned teacher at the end of the year. Teacher and student identifiers were masked by the district to preserve anonymity.

Data

During the 2002-03 academic year, the Los Angeles Unified School District (LAUSD) enrolled 746,831 students (kindergarten through grade 12) and employed 36,721 teachers in 689 schools scattered throughout Los Angeles County. There were 429 elementary schools in the district.⁴

We use test score data from the spring of 1999 through the spring of 2007. Between the spring of 1999 and the spring of 2002, the Los Angeles Unified School District administered the Stanford 9 achievement test. State regulations did not allow for exemptions for students with disabilities or poor English skills. In the Spring of 2003, the district (and the state) switched from the Stanford 9 to the California Achievement Test. Beginning in 2004, the district used a third test—the California Standards Test. For each test and each subject, we standardized by grade and year.

Although there was considerable mobility of students within the school district (9 percent of students in grades 2 through 5 attended a different school than they did the previous year), the geographic size of LAUSD ensured that most students remained within the district even if they moved. Conditional on having a baseline test score, we observed a follow-up test score for 90 percent of students in the following spring.

We observed snapshots of classroom assignments in the fall and spring semesters. In both the experimental and non-experimental samples, our analysis focuses on “intention to treat” (ITT), using the characteristics of the teacher to whom a student was assigned in the fall.

We also obtained administrative data on a range of other demographic characteristics and program participation. These included race/ethnicity (hispanic, white,

⁴ Student enrollment in LAUSD exceeds that of 29 states and the District of Columbia.

black, other or missing), indicators for those ever retained in grade, designated as Title I students, those eligible for Free or Reduced Price lunch, those designated as homeless, migrant, gifted and talented or participating in special education. We also used information on tested English language Development level (level 1-5). In many specifications, we included fixed effects for the school, year, calendar track and grade for each student.⁵

We dropped those students in classes where more than 20 percent of the students were identified as special education students. In the non-experimental sample, we dropped classrooms with extraordinarily large (more than 36) or extraordinarily small (less than 10) enrolled students. (This restriction excluded 3 percent of students with valid scores). There were no experimental classrooms with such extreme class sizes.

Empirical Methods

Our empirical analysis proceeded in two steps. In the first step, we used a variety of standard methods to estimate teacher value added based on observational data available prior to the experiment. In the second step, we evaluated whether these value-added estimates accurately predicted differences in student's end-of-year test scores between pairs of teachers who were randomly assigned to classrooms in the subsequent experimental data.

As emphasized by Rubin, Stuart and Zanutto (2004), it is important to clearly define the quantity we are trying to estimate in order to clarify the goal of value-added estimation. Our value-added measures are trying to answer a very narrow question: If a

⁵ Because of overcrowding, LAUSD operates a number of schools on a year-round calendar—with students on up to four different schedules rotating their attendance throughout the year, which we refer to a calendar track.

given classroom of students were to have teacher A rather than teacher B, how much different would their average test scores be at the end of the year? Thus, the outcome of interest is end-of-year test scores, the treatment that is being applied is the teacher assignment, and the unit at which the treatment occurs is the classroom. We only observe each classroom with its actual teacher, and do not observe the counter-factual case of how that classroom would have done with a different teacher. The empirical challenge is estimating what test scores would have been in this counter-factual case. When teachers are randomized to classrooms (as in our experimental data), classroom characteristics are independent of teacher assignment and a simple comparison of average test scores among each teacher's students is an unbiased estimate of differences in teacher value added. The key issue that value added estimates must address is the potential non-random assignment of teachers to classrooms in observational data, i.e. how to identify "similar" classrooms that can be used to estimate what test scores would have been with the assignment of a different teacher.

While there are many other questions we might like to ask – such as, “what is the effect of switching a single student across classrooms,” or “what is the effect of peer or school characteristics”, or “what is the effect on longer-run student outcomes” – these are not the goal of the typical value-added estimation. Moreover, estimates of value added tell us nothing about *why* a given teacher affects student test scores. Although we are assuming a teacher's impact is stable, it may reflect the teacher's knowledge of the material, pedagogical approach, or the way that students and their parents respond to the teacher with their own time and effort. Finally, the goal of value-added estimation is not to estimate the underlying education production function. Such knowledge is relevant to

many interesting policy questions related to how we should interpret and use value added estimates, but estimating the underlying production function requires extensive data and strong statistical assumptions (Todd and Wolpin, 2003). The goal of estimating teacher value added is much more modest, and can be accomplished under much weaker conditions.

Step 1: Estimating teacher value added with prior observational data

To estimate the value added of the teachers in our experiment, we used four years of data available prior to the experiment (1999-2000 through 2002-2003 school years). Data on each student's teacher, background characteristics, end of year tests, and prior year tests were available for students in grades 2 through 5. To make our observational sample comparable to our experimental sample, we limited our sample to the schools that participated in the experiment. To assure that our observational sample was independent of our experimental sample, we excluded all students who were subsequently in any of our experimental classrooms (e.g., 2nd graders who we randomly assigned a teacher in a later grade). We also excluded students in classrooms with fewer than five students in a tested grade, as these classrooms provided too few students to accurately estimate teacher value added (and were often a mixed classroom with primarily 1st graders). After these exclusions, our analysis sample included data on the students of 1950 teachers in the experimental schools, including 165 teachers who were later part of the experimental analysis.

Teacher value added was estimated as the teacher effect (μ) from a student-level estimating equation of the general form:

$$(1) \quad A_{ijt} = X_{ijt}\beta + v_{ijt}, \quad \text{where } v_{ijt} = \mu_j + \theta_{jt} + \varepsilon_{ijt}$$

The dependent variable (A_{ijt}) was either the end-of-year test score (standardized by grade and year) or the test score gain since the prior spring for student i taught by teacher j in year t . The control variables (X_{ijt}) included student and classroom characteristics, and are discussed in more detail below. The residual (v_{ijt}) was assumed to be composed of a teacher's value added (μ_j) that was constant for a teacher over time, an idiosyncratic classroom effect (to capture peer effects and classroom dynamics) that varied from year to year for each teacher (θ_{jt}), and an idiosyncratic student effect that varied across students and over time (ε_{ijt}).

A variety of methods have been used in the literature to estimate the coefficients (β) and teacher effects (μ) in equation 1 (see McCaffrey, 2003, for a recent survey). We estimated equation 1 by OLS, and used the student residuals (v) to form empirical Bayes estimates of each teacher's value added as described in greater detail below (Morris, 1983). If the teacher and classroom components are random effects (uncorrelated with X), OLS estimation yields consistent but inefficient estimates of β . Hierarchical Linear Models (HLM) were designed to estimate models such as equation 1 with nested random effects, and are a commonly used alternative estimation method that yields efficient maximum likelihood estimates of β at the cost of greater computational complexity (Raudenbush and Bryk, 2002). Because of our large sample sizes, HLM and OLS yield very similar coefficients and the resulting estimates of teacher value added are virtually identical (correlation > .99). Another common estimation approach is to treat the teacher and classroom effects in equation 1 as fixed effects (or correlated random effects), allowing for potential correlation between the control variables (X) and the teacher and

classroom effects (Gordon, Kane, and Staiger, 2006; Kane, Rockoff and Staiger, forthcoming; Rockoff, 2004; Rothstein, 2007). Because both methods rely heavily on the within-classroom variation to identify the coefficients on X, fixed effect and OLS also yield very similar coefficients and the resulting estimates of teacher value added are therefore also very similar in our data.

While estimates of teacher value added were fairly robust to how equation 1 was estimated, they were less robust to the choice of the dependent and independent variables. Therefore, we estimated a number of alternative specifications that, while not exhaustive, were representative of the most commonly used specifications (McCaffrey, 2003). Our first set of specifications used the end-of-year test score as the dependent variable. The simplest specification included no control variables at all, essentially estimating value added based on the average student test scores in each teacher's classes. The second specification added controls for student baseline scores from the previous spring (math, reading and language arts) interacted with grade, indicators for student demographics (gender, race, migrant, participation in gifted and talented programs, participation in the free/reduced price lunch program, and grade indicators for each year), and the means of all of these variables at the classroom level (to capture peer effects). The third specification added indicators for each school to the control variables. The fourth specification replaced the student-level variables (both demographics and baseline scores) with student fixed effects. Finally, we repeated all of these specifications using test score gains (the difference between end-of-year scores and the baseline score from the previous spring) as the dependent variable. For these specifications, we excluded baseline scores from the list of control variables, which is equivalent to imposing a

coefficient of one on the baseline score in the levels specification. Student fixed effects were highly insignificant in the gains specification, so we do not report value added estimates for this specification. Each of the specifications was estimated separately by subject, yielding seven separate value-added measures (four using test levels, three using test gains) for each teacher in math and language arts.

For each specification, we used the student residuals (v) from equation 1 to form empirical Bayes estimates of each teacher's value added (Raudenbush and Bryk, 2002). This is the approach we have used successfully in our prior work (Gordon, Kane, and Staiger, 2006; Kane, Rockoff and Staiger, forthcoming; Rockoff, 2004). The empirical Bayes estimate is a best linear predictor of the random teacher effect in equation 1 (minimizing the mean squared prediction error), and under normality assumptions is an estimate of the posterior mean (Morris, 1983). The basic idea of the empirical Bayes approach is to multiply a noisy estimate of teacher value added (e.g., the mean residual over all of a teacher's students from a value added regression) by an estimate of its reliability, where the reliability of a noisy estimate is the ratio of signal variance to signal plus noise variance. Thus, less reliable estimates are shrunk back toward the mean (zero, since the teacher estimates are normalized to be mean zero). Nearly all recent applications have used a similar approach to estimate teacher value added (McCaffrey et al., 2003).

We constructed the empirical Bayes estimate of teacher value added in three steps.

- 1) First, we estimated the variance of the teacher (μ_j), classroom (θ_{jt}) and student (ε_{ijt}) components of the residual (v_{ijt}) from equation 1. The within-classroom variance in

v_{ijt} was used as an estimate of the variance of the student component:

$$(2) \quad \hat{\sigma}_\varepsilon^2 = \text{Var}(v_{ijt} - \bar{v}_{jt}).$$

The covariance between the average residual in a teacher's class in year t and year t-1 was used as an estimate of the variance in the teacher component:⁶

$$(3) \quad \hat{\sigma}_\mu^2 = \text{Cov}(\bar{v}_{jt}, \bar{v}_{jt-1}).$$

The covariance calculation was weighted by the number of students in each classroom (n_{jt}). Finally, we estimated the variance of the classroom component as the remainder:

$$(4) \quad \hat{\sigma}_\theta^2 = \text{Var}(v_{ijt}) - \hat{\sigma}_\mu^2 - \hat{\sigma}_\varepsilon^2.$$

- 2) Second, we formed a weighted average of the average classroom residuals for each teacher (\bar{v}_{jt}) that was a minimum variance unbiased estimate of μ_j for each teacher (so that weighted average had maximum reliability). Data from each classroom was weighted by its precision (the inverse of the variance), with larger classrooms having less variance and receiving more weight:

$$(5) \quad \bar{v}_j = \sum_t w_{jt} \bar{v}_{jt}, \text{ where } w_{jt} = \frac{h_{jt}}{\sum_t h_{jt}} \text{ and}$$

$$h_{jt} = \frac{1}{\text{Var}(\bar{v}_{jt} | \mu_j)} = \frac{1}{\hat{\sigma}_\theta^2 + \left(\frac{\hat{\sigma}_\varepsilon^2}{n_{jt}} \right)}$$

⁶ This assumes that the student residuals are independent across a teacher's classrooms. Occasionally, students will have the same teacher in two subsequent years – either because of repeating a grade, or because of looping (where the teacher stays with the class through multiple grades). Since this occurs infrequently in our data, we have assumed that student residuals are uncorrelated across years. It is not difficult to allow for correlation in student residuals in equation 3, by either directly adjusting for this correlation or by dropping classrooms with common students. We plan to do this in the future.

3) Finally, we constructed an empirical Bayes estimator of each teacher's value added by multiplying the weighted average of classroom residuals (\bar{v}_j) by an estimate of its reliability:

$$(6) \quad VA_j = \bar{v}_j \left(\frac{\hat{\sigma}_\mu^2}{Var(\bar{v}_j)} \right), \text{ where } Var(\bar{v}_j) = \hat{\sigma}_\mu^2 + \left(\sum_t h_{jt} \right)^{-1}$$

The quantity in parenthesis represents the shrinkage factor, and reflects the reliability of \bar{v}_j as an estimate of μ_j , where the reliability is the ratio of signal variance to total variance. Note that the total variance is the sum of signal variance and estimation error variance, and the estimation variance for \bar{v}_j can be shown to be $\left(\sum_t h_{jt} \right)^{-1}$.

Step 2: Experimental validation of non-experimental value-added estimates

In the experimental data, we evaluated both the bias and predictive accuracy of the value-added estimates generated by each of our specifications. If teachers were randomly assigned to classrooms in the non-experimental data, then specifications with additional controls could improve the precision of the value-added estimates but would not affect bias. If teachers were not randomly assigned to classrooms in the non-experimental data, then additional controls could also reduce bias. Thus, both bias and predictive accuracy are questions of interest.

The experimental data consisted of information on all students originally assigned to 78 pairs of classrooms. As discussed below, some students and teachers changed classrooms subsequent to randomization. All of our analyses were based on the initial

teacher assignment of students at the time of randomization, and therefore represent an intention-to-treat analysis.

Since randomization was done at the classroom-pair level, the unit of our analysis was the classroom-pair (with only secondary analyses done at the student level) which provides 78 observations. For our main analyses, we averaged student-level data for each classroom, and estimated the association between these classroom-level outcomes and teacher value added. Teachers were randomized within but not across pairs, so our analysis focused on within-pair differences, and estimated models of the form:

$$(7) \quad \bar{Y}_{jp} = \alpha_p + \beta VA_{jp} + \varepsilon_{jp}, \text{ for } j=1,2 \text{ and } p=1,\dots,78.$$

The dependent variable is an average outcome for students assigned to the classroom, the independent variable is the assigned teacher's value-added estimate, and we control for pair fixed effects. Since there are two classrooms per randomized group, we estimated the model in first differences (which eliminates the constant, since the order of the teachers is arbitrary):

$$(8) \quad \bar{Y}_{2p} - \bar{Y}_{1p} = \beta(VA_{2p} - VA_{1p}) + \tilde{\varepsilon}_p, \text{ for } p=1,\dots,78.$$

These bivariate regressions were run un-weighted, and robust standard errors were used to allow for heteroskedasticity across the classroom pairs. In secondary analyses we estimated equation 7 at the student level (which implicitly weights each class by the number of students) and clustered the standard errors at the pair level.

To validate the non-experimental value-added estimates, we estimated equation 8 using the within-pair difference in end-of-year test scores (math or language arts) as the dependent variable. A coefficient of one on the difference in teacher value added would indicate that the value added measure being evaluated was unbiased – that is, the

expected difference between classrooms in end-of-year tests scores is equal to the difference between the teachers' value added. In fact, we might expect a coefficient somewhat different from one because our intention-to-treat analysis is based on initial assignment, while about 15 percent of students have a different teacher by the time of the spring test. We use the R-squared from these regressions to evaluate the predictive accuracy of each of our value-added measures.

We also explore the persistence of teacher effects on test scores by estimating equation 8 using differences in student achievement one and two years after the experimental assignment to a particular teacher. McCaffrey et al. (2004) found that teacher effects on math scores faded out in a small sample of students from five elementary schools. In our experimental setting, the effect of teacher value added on student achievement one or two years later could be the result of this type of fade out or could be the result of systematic teacher assignment in the years subsequent to the experiment. We report some student-level analyses that control for subsequent teacher assignment (comparing students randomly assigned to different teachers who subsequently had the same teacher), but these results are no longer purely experimental since they condition on actions taken subsequent to the experiment.

Finally, we estimate parallel regressions based on equation 8 to test whether baseline classroom characteristics or student attrition are related to teacher assignment. Using average baseline characteristics of students in each class as the dependent variable, we test whether teacher assignment was independent of classroom and student characteristics. Similarly, using the proportion of students in each class who were missing the end-of-year test score as the dependent variable, we test whether student attrition was related to teacher assignment. We expect a coefficient of zero on the difference in

teacher value added in these regressions, implying that classroom characteristics and student attrition were not related to teacher assignment. While only 10% of students are missing end-of-year test scores, selective attrition related to teacher assignment is a potentially serious threat to the validity of our experiment.

Sample Comparisons

Table 1 reports the characteristics of three different samples. An “experimental school” is any school which contained a pair of classrooms that was included in the random assignment experiment. Within the experimental schools, we have reported separately the characteristics of teachers in the experimental sample and those that were not. The teachers in the experimental sample were somewhat less likely to be Hispanic than the other teachers in the experimental schools (26 percent versus 31 percent) and somewhat more likely to be African American (17 percent versus 14 percent). The average experimental teacher also had considerably more teaching experience, 15.5 years versus 10.5 years. Both of these differences were largely due to the sample design, which focused on applicants to the National Board for Professional Teaching Standards.

We also used the full sample of students, in experimental and non-experimental schools, to estimate non-experimental teacher effects conditioning on student/peer characteristics and baseline scores. Although the teachers in the experimental sample differed from those in the non-experimental sample in some observable characteristics, the mean and standard deviation of the non-experimental teacher effects were very similar across the three samples.

In Table 2, we compare student characteristics across the same three groups, including mean student scores in 2004 through 2007 for students in the experimental

schools and non-experimental schools. Although the racial/ethnic distributions are similar, three differences are evident. First, within the experimental schools, the students assigned to the experimental sample of teachers had somewhat higher test scores, .027 standard deviations above the average for their grade and year in math, while the non-experimental sample had baseline scores .11 standard deviations below the average. We believe this too is a result of the focus on National Board applicants in the sample design, since more experienced teachers tend to be assigned students with higher baseline scores. Second, the student baseline scores in the non-experimental schools are about .024 standard deviations higher than average. Third, the students in the experimental sample are more likely to be in 2nd and 3rd grade, rather than 4th and 5th grade. Again, this is a result of the sample design: in Los Angeles, more experienced teachers tend to concentrate in grades K-3, which have small class sizes (20 or fewer students) as a result of the California class size reduction legislation.

Estimates of Variance Components of Teacher Effects

Table 3 reports the various estimates that were required for generating our empirical Bayes estimates of teacher effects. The first column reports the estimate of the standard deviation in “true” teacher impacts. Given that students during the pre-experimental period were generally not randomly assigned to classrooms, our estimate of the standard deviation in true teacher effects is highly sensitive to the student-level covariates we use. For instance, if we include no student-level or classroom-level mean baseline characteristics as covariates, we would infer that the standard deviation in teacher impacts was .455 in math and .458 in English language arts. However, after including covariates for student and peer baseline performance and characteristics, the implied s.d. in teacher effects is essentially cut in half, to .228 in math and .182 in English language arts. Adding controls for school effects has little

impact, lowering the estimated s.d. in teacher impacts to .216 in math and .173 in English language arts. (Consistent with earlier findings, this reflects the fact that the bulk of the variation in estimated teacher effects is among teachers working in the same school, as opposed to differences in mean estimated impact across schools.) However, adding student by school fixed effects, substantially lowers the estimated s.d. in teacher impact to .098 and .082.

A standard deviation in teacher impact in the range of .18 to .20 is quite large. Since the underlying data are standardized at the student and grade level, an estimate of that magnitude would imply that the difference between being assigned a 25th or a 75th percentile teacher would imply that the average student would improve about one-quarter of a standard deviation relative to similar students in a single year.

The second column reports our estimate of the standard deviation of the classroom by year error term. These errors—which represent classroom-level disturbances such as a dog barking on the day of the test or a coincidental match between a teacher’s examples and the specific questions that appeared on the test that year-- are assumed to be i.i.d. for each teacher for each year. Rather than being trivial, this source of error is estimated to be quite substantial and nearly equal to the standard deviation in the signal (e.g. a standard deviation of .180 for the classroom by year error term in math versus .228 for the estimated teacher impact on math after including student and peer-level covariates) In English language arts, the estimated standard deviation in the teacher signal is essentially equal to the standard deviation in the classroom by year error.

The third component in the table is the standard deviation in the within-classroom- and year student-level error term. And the fourth column in the table reports the mean number of observations we had for each teacher (summed across years) for estimating their

effect. Across the 4 school years (spring 2000 through spring 2003), we observed an average of 42 to 50 student scores per teacher for estimating teacher effects.

Relationship between Pre-experimental Estimates and Baseline Characteristics

To the extent that classrooms were randomly assigned to teachers, we would not expect a relationship between teacher's non-experimental value-added estimates and the characteristics of their students. Indeed, as reported in Table 4, there is no significant relationship between the within-pair difference in pre-experimental estimates of teacher effects and the differences in student performance or characteristics (baseline math and reading, participation in the gifted and talented program, Title I, the free or reduced price lunch program, race/ethnicity, an indicator for those students retained in a prior grade).⁷

Attrition and Teacher Switching

In Table 5, we report relationships between the within-pair difference in pre-experimental estimates of teacher effects and the difference in proportion of students missing test scores at the first, second or third year following random assignment. For the entry in the first row of column (1), we estimated the relationship between the within-pair difference in pre-experimental teacher math effects and the difference in the proportion of students missing math scores at the end of the first year. Analogously, the second row reports the relationship between within-pair differences in pre-experimental ELA effects and the proportion missing ELA scores. There is no statistically significant relationship between

⁷ Since random assignment occurred at the classroom level (not the student level), we take the first-difference within each pair and estimate each of these relationships with one observation per pair. In results not reported here, we also explored the relationship using student-level regressions, including fixed effects for each pair and clustering at the pair level. None of those relationships were statistically significant either.

pre-experimental teacher effect estimates and the proportion missing test scores in the first, second or third year. Thus, systematic attrition does not appear to be a problem.

The last column reports the relationship between pre-experimental value-added estimates for teachers and the proportion of students switching teachers during the year. Although about 15 percent of students had a different teacher at the time of testing than they did in the fall semester, there was no relationship between teacher switching and pre-experimental value-added estimates.

Experimental Outcomes

Table 6 reports the relationship between within-pair differences in mean test scores for students at the end of the experimental year (as well as for the subsequent two years when students are dispersed to other teachers' classes) and the within-pair differences in pre-experimental teacher effects. The pre-experimental teacher effects were estimated using a variety of specifications. We first estimated teacher effects using test score levels with no covariates. We added a variety of controls incrementally-- starting with student- and classroom-level controls for baseline characteristics and test scores, then adding school fixed effects, then adding fixed effects by student and school.

Rather than estimating a coefficient on prior score, a number of researchers have used test score gains as the dependent variable (score in year t minus score in year $t-1$)—effectively constraining the coefficient on baseline scores to be equal to one. Accordingly, we also estimated teacher effects using gains but no other covariates, and successively added student and peer controls (not including baseline test scores) and school fixed effects.

The coefficients on the within-pair difference in each of these pre-experimental measures of teacher effects in predicting the within-pair difference in the mean of the

corresponding test score (whether math or English language arts) are reported in Table 7. Each of these was estimated with a separate bivariate regression with no constant term. Several findings are worth noting.

First, all of the coefficients on the pre-experimental estimates in column (1) are statistically different from zero. Whether using test score levels or gains, or math or English language arts, those classrooms assigned to teachers with higher non-experimental estimates of effectiveness scored higher on both math and English language arts at the end of the first school year following random assignment.

Second, those pre-experimental teacher effects that fail to control for any student or peer-level covariates are biased. Recall from the discussion in the empirical methods section, each of the estimated teacher effects have been “shrunk” to account for random sources of measurement error—both non-persistent variation in classroom performance and student-level errors. If there were no bias, we would expect the coefficients on the adjusted pre-experimental estimate of teacher effects to be equal to one. Although we could reject the hypothesis that mean student scores in years prior to the experiment had a coefficient of zero, we could also reject the hypothesis that the coefficients were equal to one: in math, the 95 percent confidence interval was $.495 \pm 1.96 * .103$, while the confidence interval in ELA was $.377 \pm 1.96 * .149$. The fact that the coefficient is less than one implies that a 1-point difference in prior estimated value-added is associated with *less than* 1 point (in fact, about half that) difference in student achievement at the end of the year. To the extent that students were *not* randomly assigned to teachers during the pre-experimental period, we would have expected the pre-experimental estimates using test score levels to have been biased upward in this way if better teachers were being assigned students with higher baseline achievement.

Third, the coefficients on the pre-experimental teacher effects which used student-level fixed effects were close to 2 (1.987 in math and 2.35 in English language arts) and the 95 percent confidence interval does not include one. Apparently, such estimates tend to understate true variation in teacher effects. As Rothstein (2007) has argued, one might expect that the specification with student- and teacher-fixed effects would be biased toward understating the variation in teacher effects, to the extent that the student effects partially *absorb* the effect of teachers on subsequent student performance.

Fourth, note that the coefficients on the estimated teacher effects in the remaining specifications (test score levels with student and peer controls, or test score gains with or without including other student and peer controls) were all close to 1, significantly greater than zero, and not statistically different from one. In other words, we could reject the hypothesis that they had no relationship to student performance, but we could not reject the hypothesis that the pre-experimental estimates of teacher effects were unbiased.

Fifth, in terms of being able to predict differences in student achievement at the end of the experimental year, the specifications using pre-experimental estimates based on student/peer controls and school fixed effects had a slightly higher R^2 -- .230 for math and .163 in English language arts. In other words, of the several specifications which we could not reject as being unbiased, the specification with the lowest mean squared error in terms of predicting differences in student achievement within schools was that which included student/peer controls and school fixed effects. (Recall that the experimental design is also focused on measuring differences in student achievement within schools, so those too implicitly include school fixed effects.)

To illustrate this predictive ability of pre-experimental estimates, we plotted the difference in student achievement within teacher pairs against the difference in pre-

experimental teacher effects for these preferred specifications in Figure 1 (math) and 2 (English language arts), along with the estimated regression line and the prediction from a lowess regression. Teachers were ordered within the randomized pair so that the values on the x-axis are positive, representing the difference between the higher and lower value-added teacher. Thus, we expect the difference in achievement between the two classrooms to be positive, and more positive as the difference in value-added increases between the two teachers. This pattern is quite apparent in the data, and both the regression line and the lowess predictions lie near to the 45 degree line as expected.

Finally, the remaining columns of Table 6 report differences in student achievement one and two years after the experimental assignment to a particular teacher. After students have dispersed into other teachers' classrooms in the year following the experiment, about half of the math impact had faded. (Each of the coefficients declines by roughly 50 percent.) In the second year after the experimental year, the coefficients on the teacher effects on math had declined further and were not statistically different from zero. In other words, while the mean student assigned to a high "value-added" teacher seems to outperform similar students at the end of the year, the effects fade over the subsequent two years. As discussed in the conclusion, this has potentially important implications for calculating the cumulative impact of teacher quality on math achievement.

However, the degree of fade-out in English language arts is much less pronounced. Even in the second year following the experimental year, the effect of having been assigned a high "value-added" teacher remains statistically different from zero.

Testing for Compensatory Teacher Assignment

If principals were to compensate a student for having been assigned a high- (or low-) value-added teacher one year with a low (or high-) value-added teacher the next year, we would be overstating the degree of fade-out in the specifications above. That is, a student randomly assigned a high-impact teacher during the experiment might have been assigned a low-impact teacher the year after. Although the strategy relies on quasi-experimental identification and not random assignment one way to test this hypothesis is to re-estimate the relationships using student-level data and include fixed effects for teacher assignments in subsequent years. As reported in Table 7, there is little reason to believe that compensatory teacher assignments accounts for the fade-out in math. The first two columns report results from student-level regressions that were similar to the pair-level regression reported for first and second year scores in the previous table. The only difference from the corresponding estimates in Table 6 is that these estimates are estimated at the student level and, therefore, place larger weight on classrooms with more students. As we would have expected under random assignment, adding student-level covariates has no appreciable effect on the estimated coefficient on prior teacher effects. The third column of Table 7 reports the coefficient on one's experimental year teacher in predicting one's subsequent performance, including fixed effects for one's teacher in the subsequent year. Sample size falls somewhat in these regressions because we do not have reliable teacher assignments for students who were subsequently in middle school grades. If principals were assigning teachers in successive years to compensate (or to ensure that students have similar mean teacher quality over their stay in school), one would expect the coefficient on the experimental year teacher's effect to rise once the teacher effects are added. The coefficient is little changed. The same is true in the second year after the experimental year.

Fade-Out in the Non-Experimental Data

Given the extent of the fade-out in the experimental sample, we investigated the degree of fade-out for the non-experimental sample using an analogous framework. First, we generated an estimate of each teacher's impact on math and English language arts scores, pooling data from the pre-experimental period. We limited the sample to students and teachers in schools that did *not* have any teachers in the experimental sample. We then tested the predictive power of the pre-experimental estimates—estimated for earlier cohorts of children—in predicting student performance for students in grades 3 through 5 in 2004-05. (By 2004-05, the district had discontinued testing at the end of grade 1, so there were no baseline scores for students in grade 2.)

The results are reported in Table 8. As reported in column (1), the coefficient on a teacher's prior value-added was 1.096 in math and .87 in language arts. In column (2), we add covariates with student and peer baseline scores and characteristics. Given that these are observational data, we might expect some relationship between student baseline characteristics and teacher effects. The coefficients on math and English language arts decline somewhat to .95 and .75 respectively.

In column (3), we use test performance at the end of the 2005-06 school year as the dependent variable, although we continue to condition on baseline performance from the spring of 2004 and use the value-added estimate for their 2004-05 teacher as a regressor. The results imply even more fade-out than observed in the experimental sample. The coefficient on both math and language arts declined to .246 and .223 respectively. The last two columns use performance in 2006-07 as the outcome. The impact of the 2004-05 teacher two years later was only about one tenth as large (.115 and .140 in math and language

arts respectively) as the one-year impact estimated during the pre-experimental period. When we control for subsequent teacher assignment with fixed effects in columns (4) and (6), the fade out becomes even more pronounced.

Conclusion

Given the ubiquity of non-experimental impact evaluation in education, there is a desperate need to validate the implied causal effects with experimental data. In this paper, we have focused on measuring the extent of *bias* in non-experimental estimates of teacher effects. We could not reject the hypothesis that several common non-experimental specifications yielded unbiased estimates. However, such work needs to be replicated in other districts before a consensus is built around the validity of such non-experimental estimates. After all, there may be something idiosyncratic about the assignment of teachers to students in Los Angeles. For instance, in Los Angeles, as in many urban districts, parents may be less involved in advocating for specific teachers for their children than in suburban districts. There may also be less tracking and collective bargaining agreements may make it harder for principals to play favorites in assigning rosters to teachers.

Several recent papers, such as Rothstein (2007) have used non-experimental data to test the strong exclusion restrictions typically used in value-added models, such as the exclusion of prior and subsequent teacher effects. As he has reported, some of those exclusion restrictions are rejected by the data. That finding is not inconsistent with our results. For instance, both our experimental and non-experimental analyses above imply that there may be significant fade-out of teacher effects from one year to the next, particularly in math. Such fade-out would imply that a prior teacher's effect would enter into subsequent year performance. However, excluding the effects of prior teachers would create bias only to the extent that teacher quality is correlated over time. Even in our non-experimental data,

there was little evidence that the quality of one's teacher from year to year was correlated.⁸ Teacher assignments from year to year are neither compensatory nor magnifying teacher effects in prior years.

The degree of fade-out in math and reading effects would have strong implications for a wide array of policy analyses using educational impact estimates. (Interestingly, McCaffrey et al. (2004) report considerable fade-out in non-experimental data, although the effect is very imprecisely estimated.⁹) When calculating the potential value of shifting the teacher effectiveness distribution, we and others have typically assumed that the effects of a strong teacher persist in the children they teach. Our results call that assumption into question, particularly for math. While impacts on reading and language arts skills tend to "stick", impacts on math seemed to fade if not reinforced.

⁸ The same seemed to be true in Rothstein's data, since the estimated effects of one's current teacher with the exclusion restrictions in place were highly correlated with the estimates when those restrictions were lifted. In fact, in Table 12, the reported correlation was as high as .99.

⁹ McCaffrey et al. (2004), Table 1, p. 90. Their estimates imply that only 20 percent of a teacher's effect persists between 3rd and 4th grade and 30 percent of the effect of a 4th grade teacher persists through grade 5. However, the standard error on both estimates is .20.

References:

- Aaronson, Daniel, Lisa Barrow and William Sander (2007) "Teachers and Student Achievement in Chicago Public High Schools" *Journal of Labor Economics* Vol. 24, No. 1, pp. 95-135.
- Andrabi, Tahir, Jishnu Das, Asim I. Khwaja, Tristan Zajonc (2008) "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics" Harvard University unpublished working paper, Feb. 19.
- Armour, David. T. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools*. R-2007-LAUDS. (Santa Monica, CA: Rand Corporation).
- Gordon, Robert, Thomas J. Kane and Douglas O. Staiger, (2006) "Identifying Effective Teachers Using Performance on the Job" Hamilton Project Discussion Paper, Published by the Brookings Institution.
- Hanushek, Eric A. (1971). "Teacher characteristics and gains in student achievement; estimation using micro data". *American Economic Review*, 61, 280-288.
- Jacob, Brian and Lars Lefgren (2005) "Principals as Agents: Student Performance Measurement in Education" *NBER Working Paper No. 11463*.
- Kane, Thomas J., Jonah Rockoff and Douglas Staiger, (Forthcoming) "What Does Certification Tell Us about Teacher Effectiveness?: Evidence from New York City" *Economics of Education Review* (Also NBER Working Paper No. 12155, April 2006.
- McCaffrey, Daniel, J.R. Lockwood, Daniel Koretz and Laura Hamilton (2003) *Evaluating Value-Added Models for Teacher Accountability*, (Santa Monica, CA: Rand Corporation).
- McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, Laura Hamilton (2004) "Models for Value-Added Modeling of Teacher Effects" *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, Value-Added Assessment Special Issue., Spring, pp. 67-101.
- Morris ,Carl N (1983) "Parametric Empirical Bayes Inference: Theory and Applications" *Journal of the American Statistical Association*, 78:47-55.
- Murnane, R. J. & Phillips, B. R. (1981). "What do effective teachers of inner-city children have in common?" *Social Science Research*, 10, 83-100.
- Nye, Barbara, Larry Hedges and Spyros Konstantopoulos (2004) "How large are teacher effects?" *Educational Evaluation and Policy Analysis* Volume 26, pp. 237-257.

- Raudenbush, Stephen W. (2004) "What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?" *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, Value-Added Assessment.Special Issue. Spring, pp. 121-129.
- Raudenbush, Stephen W. and A.S. Bryk (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Newbury Park, CA: Sage Publications.
- Rivkin, Steven, Eric Hanushek and John Kain (2005) "Teachers, Schools and Academic Achievement" *Econometrica* Vol. 73, No. 2, pp. 417-458.
- Rockoff, Jonah E. (2004) "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data" *American Economic Review* Vol. 92, No. 2, pp. 247-252.
- Rubin, Donald B., Elizabeth A. Stuart; Elaine L. Zanutto (2004) "A Potential Outcomes View of Value-Added Assessment in Education" *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, Value-Added Assessment Special Issue, Spring, pp. 103-116.
- Sanders, William L. and June C. Rivers (1996) "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement" *Research Progress Report* University of Tennessee Value-Added Research and Assessment Center.
- Todd, Petra E. and Kenneth I. Wolpin (2003) "On the Specification and Estimation of the Production Function for Cognitive Achievement" *Economic Journal* Vol. 113, No. 485.

Figure 1: Within Pair Differences in Pre-experimental Value-added and End of First Year Test Score
Mathematics

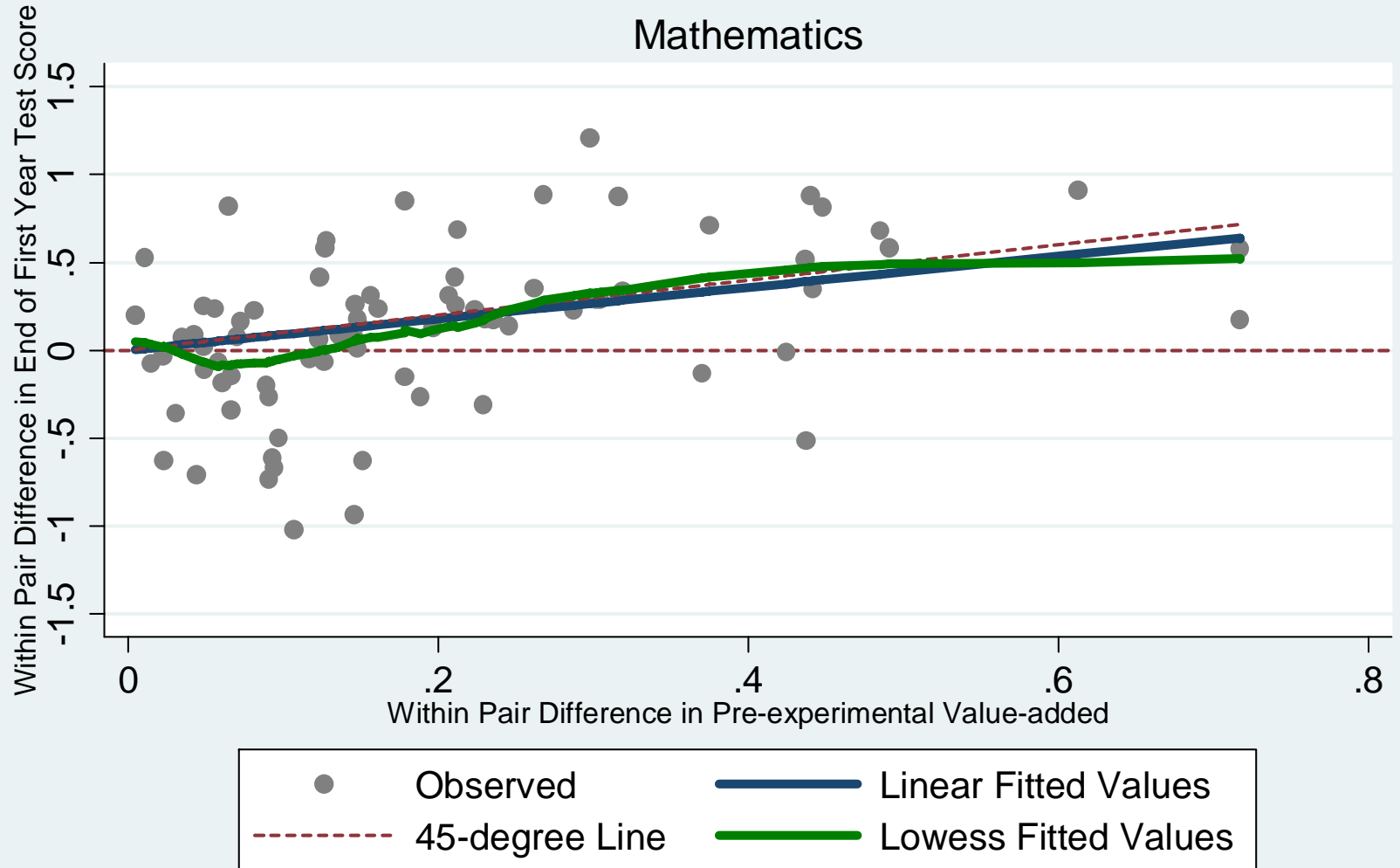


Figure 2: Within Pair Differences in Pre-experimental Value-added and End of First Year Test Score

English Language Arts

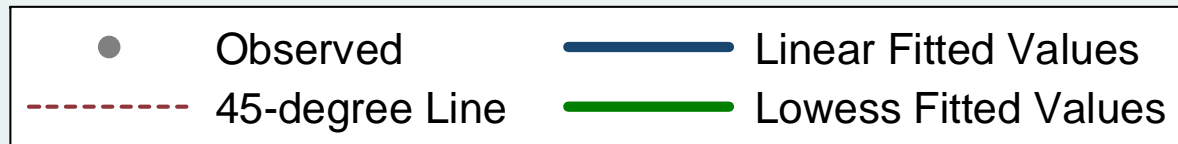
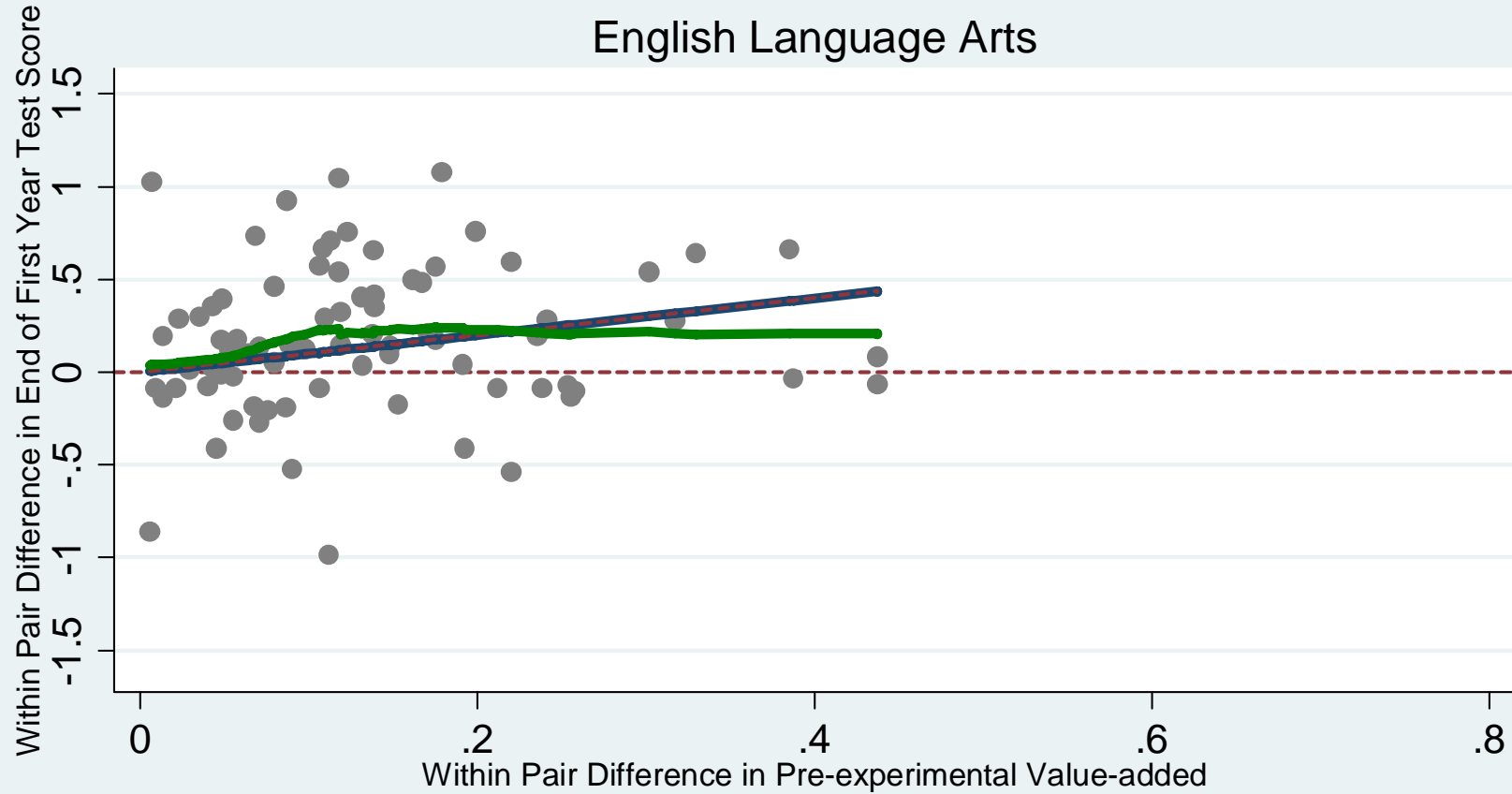


Table 1: Sample Comparison - Teachers

	Experimental School		Non-experimental School
	Experimental Sample	Non-experimental Sample	Non-experimental Sample
Mean Teacher Effect in Math	-0.009	-0.003	0.005
S.D.	0.195	0.196	0.196
Mean Teacher Effect in ELA	-0.010	0.003	0.003
S.D.	0.149	0.148	0.147
Black, Non-Hispanic	0.166	0.138	0.123
Hispanic	0.258	0.311	0.325
White, Non-Hispanic	0.466	0.447	0.425
Other, Non-Hispanic	0.110	0.102	0.123
Teacher Race/Ethnicity Missing	0.000	0.003	0.003
Years of Experience	15.490	10.542	10.758
N:	165	1,785	11,352

Note: Descriptive statistics based on the experimental years (2003-04 and 2004-05). The mean teacher effect in math and ELA were estimated using the full sample of schools and teachers, controlling for baseline scores, student characteristics, and peer controls.

Table 2: Sample Comparison - Students

	Experimental School		Non-experimental School
	Experimental Sample	Non-experimental Sample	Non-experimental Sample
Math Scores			
2004 Mean	0.027	-0.110	0.024
S.D.	0.931	0.941	1.008
2005 Mean	-0.008	-0.113	0.028
S.D.	0.936	0.940	1.007
2006 Mean	0.001	-0.100	0.037
S.D.	0.960	0.941	1.006
2007 Mean	-0.016	-0.092	0.030
S.D.	0.956	0.941	1.006
ELA Scores			
2004 Mean	0.038	-0.113	0.023
S.D.	0.913	0.936	1.008
2005 Mean	0.009	-0.117	0.027
S.D.	0.920	0.930	1.009
2006 Mean	0.039	-0.096	0.037
S.D.	0.923	0.928	1.001
2007 Mean	0.018	-0.095	0.037
S.D.	0.940	0.936	1.000
Black, Non-Hispanic	0.112	0.115	0.113
Hispanic	0.768	0.779	0.734
White, Non-Hispanic	0.077	0.060	0.088
Other, Non-Hispanic	0.044	0.046	0.066
Grade 2	0.377	0.280	0.288
Grade 3	0.336	0.201	0.207
Grade 4	0.113	0.215	0.211
Grade 5	0.131	0.305	0.294
N:	3,554	43,766	273,525

Note: Descriptive statistics based on the experimental years (2003-04 and 2004-05). Students present both years are counted only once.

Table 3: Non-experimental Specifications of Teacher Effects

Specification Used for Non-experimental Teacher Effect	Standard Deviation of Each Component (in Student-level Standard Deviation Units)		Mean Sample Size per Teacher
	Teacher Effects	Teacher by Year Random Effect	
Math Levels with...			
<i>No Controls</i>	0.455	0.224	48.612
<i>Student/Peer Controls (incl. prior scores)</i>	0.228	0.180	42.843
<i>Student/Peer Controls (incl. prior scores) & School F.E.</i>	0.216	0.178	42.843
<i>Student Fixed Effects</i>	0.098	0.072	48.612
Math Gains with...			
<i>No Controls</i>	0.228	0.223	45.171
<i>Student/Peer Controls</i>	0.227	0.220	45.171
<i>Student/Peer Controls & School F.E.</i>	0.217	0.221	45.171
English Language Arts Levels with...			
<i>No Controls</i>	0.458	0.220	48.391
<i>Student/Peer Controls (incl. prior scores)</i>	0.182	0.169	42.730
<i>Student/Peer Controls (incl. prior scores) & School F.E.</i>	0.173	0.168	42.730
<i>Student Fixed Effects</i>	0.082	0.041	48.391
English Language Arts Gains with...			
<i>No Controls</i>	0.186	0.205	44.366
<i>Student/Peer Controls</i>	0.177	0.202	44.366
<i>Student/Peer Controls & School F.E.</i>	0.170	0.202	44.366

Note: The above estimates are based on the total variance in estimated teacher fixed effects using observations from the pre-experimental data (years 1999-2000 through 2002-03). See the text for discussion of the estimation of the decomposition into teacher by year random effects, student-level error, and "actual" teacher effects. The sample was limited to schools with teachers in the experimental sample. Any individual students who were in the experiment were dropped from the pre-experimental estimation, to avoid any spurious relationship due to regression to the mean, etc.

Table 4. Baseline Student Characteristics Regressed on Non-Experimental Teacher Effects

Specification Used for Non-experimental Teacher Effect	Baseline Scores		Baseline Demographics & Program Participation							English Language Status
	Math Score	Language Score	Gifted and Talented	Ever Retained	Special Education	Hispanic	Black	Title I	Free Lunch	Level 1 to 3
Math Levels with Student/Peer Controls	-0.081 (0.230)	0.036 (0.268)	-0.014 (0.022)	-0.042 (0.039)	-0.049 (0.033)	-0.053 (0.041)	0.008 (0.041)	0.037 (0.053)	0.031 (0.061)	-0.026 (0.071)
N:	44	44	78	78	78	78	78	78	78	78
ELA Levels with Student/Peer Controls	0.089 (0.323)	0.296 (0.359)	0.023 (0.032)	-0.034 (0.053)	-0.066 (0.051)	-0.037 (0.097)	0.008 (0.066)	0.092 (0.085)	0.084 (0.085)	-0.097 (0.132)
N:	44	44	78	78	78	78	78	78	78	78

Note: Each baseline characteristic listed in the columns was used as a dependent variable, regressing the within-pair difference in mean baseline characteristic on different non-experimental estimates of teacher effects. The coefficients were estimated in *separate* bivariate regressions with no constant. Robust standard errors are reported in parentheses. Baseline math and language arts scores were missing for the pairs that were in second grade.

Table 5: Attrition and Teacher Switching

Specification Used for Non-experimental Teacher Effect	Missing Test Score			Switched Teacher
	First Year	Second Year	Third Year	
Math Levels with Student/Peer Controls	-0.004 (0.049)	0.029 (0.057)	-0.018 (0.058)	-0.028 (0.133)
N:	78	78	78	78
ELA Levels with Student/Peer Controls	-0.035 (0.077)	0.006 (0.084)	0.030 (0.097)	-0.148 (0.171)
N:	78	78	78	78

Note: Each baseline characteristic listed in the columns was used as a dependent variable, regressing the within-pair difference in mean baseline characteristic on different non-experimental estimates of teacher effects. The coefficients were estimated in separate bivariate regressions with no constant. Robust standard errors are reported in parentheses.

Table 6: Predicting Math Outcomes During Experimental Period (Pair-level Regressions)

Specification Used for Non-experimental Teacher Effect	Test Score First Year			Test Score Second Year	Test Score Third Year
	Coefficient	R2	N:	Coefficient	Coefficient
Math Levels with...					
<i>No Controls</i>	0.495*** (0.103)	0.183	78	0.273** (0.103)	0.128 (0.097)
<i>Student/Peer Controls (incl. prior scores)</i>	0.863*** (0.178)	0.213	78	0.378* (0.174)	0.068 (0.137)
<i>Student/Peer Controls (incl. prior scores) & School F.E.</i>	0.918*** (0.180)	0.230	78	0.410* (0.178)	0.102 (0.140)
<i>Student Fixed Effects</i>	1.987*** (0.488)	0.161	78	0.915 (0.471)	0.344 (0.428)
Math Gains with...					
<i>No Controls</i>	0.833*** (0.204)	0.168	78	0.373 (0.191)	0.054 (0.153)
<i>Student/Peer Controls</i>	0.841*** (0.211)	0.170	78	0.381 (0.201)	0.054 (0.159)
<i>Student/Peer Controls & School F.E.</i>	0.878*** (0.217)	0.176	78	0.405 (0.210)	0.067 (0.165)

Note: Each baseline characteristic listed in the columns was used as a dependent variable, regressing the within-pair difference in mean baseline characteristic on different non-experimental estimates of teacher effects. The coefficients were estimated in separate bivariate regressions with no constant. Robust standard errors are reported in parentheses.

Table 6 (cont.): Predicting ELA Outcomes During Experimental Period (Pair-level Regressions)

Specification Used for Non-experimental Teacher Effect	Test Score First Year			Test Score Second Year	Test Score Third Year
	Coefficient	R2	N:	Coefficient	Coefficient
English Language Arts Levels with...					
<i>No Controls</i>	0.377* (0.149)	0.094	78	0.303 (0.165)	0.23 (0.149)
<i>Student/Peer Controls (incl. prior scores)</i>	0.994*** (0.263)	0.146	78	0.531 (0.280)	0.510* (0.247)
<i>Student/Peer Controls (incl. prior scores) & School F.E.</i>	1.090*** (0.274)	0.163	78	0.624* (0.299)	0.575* (0.260)
<i>Student Fixed Effects</i>	2.354*** (0.641)	0.128	78	1.533 (0.795)	1.462* (0.648)
English Language Arts Gains with...					
<i>No Controls</i>	0.780** (0.244)	0.091	78	0.234 (0.239)	0.284 (0.231)
<i>Student/Peer Controls</i>	0.856** (0.267)	0.102	78	0.332 (0.261)	0.363 (0.246)
<i>Student/Peer Controls & School F.E.</i>	0.915** (0.281)	0.108	78	0.368 (0.280)	0.389 (0.259)

Note: Each baseline characteristic listed in the columns was used as a dependent variable, regressing the within-pair difference in mean baseline characteristic on different non-experimental estimates of teacher effects. The coefficients were estimated in separate bivariate regressions with no constant. Robust standard errors are reported in parentheses.

Table 7: Predicting Experimental Performance in Math and ELA (Student-level Regressions)

Specification Used for Non-experimental Teacher Effect	First Year Score	Second Year Score		Third Year Score	
Math Levels with Student/Peer Controls	0.845*** (0.181)	0.423* (0.178)	0.421* (0.185)	0.08 (0.145)	0.076 (0.290)
N:	2,905	2,685	2,305	2,504	1,892
ELA Levels with Student/Peer Controls	1.073*** (0.271)	0.605* (0.275)	0.718* (0.280)	0.589* (0.249)	0.626 (0.376)
N:	2,903	2,691	2,312	2,503	1,891
Student-Level Controls	No	No	No	No	No
Second Year Teacher F.E.			Yes		
Second x Third Year Teacher F.E.					Yes

Note: The above were estimated with student-level regressions using fixed effects for each experimental teacher pair. Robust standard errors (in parentheses) allow for clustering at the teacher-pair level. The sample for specifications including teacher fixed effects are limited to students in grades 3-5 as teacher identifiers for secondary grades are not yet available.

Table 8. Estimating Fade-Out in the Non-Experimental Sample (Student-level Regressions)

Specification Used for Non-experimental Teacher Effect	2004-05		2005-06		2006-07	
Math Levels with Student/Peer Controls	1.096*** (0.016)	0.952*** (0.010)	0.246*** (0.011)	0.144*** (0.016)	0.115*** (0.012)	0.008 (0.026)
N:	114,767	108,505	97,908	67,079	88,993	32,429
ELA Levels with Student/Peer Controls	0.869*** (0.022)	0.745*** (0.012)	0.223*** (0.013)	0.135*** (0.020)	0.140*** (0.015)	0.067* (0.032)
N:	114,963	108,656	98,009	67,140	89,028	32,442
Student-Level Controls	No	Yes	Yes	Yes	Yes	Yes
Second Year Teacher F.E.				Yes		
Second x Third Year Teacher F.E.						Yes

Note: The 2004-05 teacher effect is estimated using data from 1999-2000 through 2002-03 excluding schools who participated in the experiment. Above we report the coefficients on that estimated 2004-05 teacher effect in predicting a student's 2004-05, 2005-06, and 2006-07 scores respectively. The sample for specifications including teacher fixed effects are limited to students in grades 3-5 as teacher identifiers for secondary grades are not yet available.

Appendix Table A: Non-experimental Specifications of Teacher Effects (Entire Sample)

Specification Used for Non-experimental Teacher Effect	Standard Deviation of Each Component (in Student-level Standard Deviation Units)		Mean Sample Size per Teacher
	Teacher Effects	Teacher by Year Random Effect	
Math Levels with...			
<i>No Controls</i>	0.529	0.225	51.326
<i>Student/Peer Controls (incl. prior scores)</i>	0.232	0.185	45.423
<i>Student/Peer Controls (incl. prior scores) & School F.E.</i>	0.222	0.185	45.061
<i>Student Fixed Effects</i>	0.104	0.078	50.931
Math Gains with...			
<i>No Controls</i>	0.232	0.232	47.747
<i>Student/Peer Controls</i>	0.230	0.229	47.747
<i>Student/Peer Controls & School F.E.</i>	0.223	0.229	47.371
English Language Arts Levels with...			
<i>No Controls</i>	0.527	0.212	51.094
<i>Student/Peer Controls (incl. prior scores)</i>	0.184	0.163	45.297
<i>Student/Peer Controls (incl. prior scores) & School F.E.</i>	0.175	0.162	44.941
<i>Student Fixed Effects</i>	0.081	0.038	50.706
English Language Arts Gains with...			
<i>No Controls</i>	0.189	0.199	46.938
<i>Student/Peer Controls</i>	0.181	0.196	46.938
<i>Student/Peer Controls & School F.E.</i>	0.175	0.196	46.571

Note: The above estimates are based on the total variance in estimated teacher fixed effects using observations from the pre-experimental data (years 1999-2000 through 2002-03). See the text for discussion of the estimation of the decomposition into teacher by year random effects, student-level error, and "actual" teacher effects. The sample includes all schools and teachers. However, any individual students who were in the experiment were dropped from the pre-experimental estimation, to avoid any spurious relationship due to regression to the mean, etc.

Appendix Table A: Non-experimental Specifications of Teacher Effects (Non-experimental Sample)

Specification Used for Non-experimental Teacher Effect	Standard Deviation of Each Component (in Student-level Standard Deviation Units)		Mean Sample Size per Teacher
	Teacher Effects	Teacher by Year Random Effect	
Math Levels with...			
<i>No Controls</i>	0.543	0.225	50.558
<i>Student/Peer Controls (incl. prior scores)</i>	0.233	0.187	44.762
<i>Student/Peer Controls (incl. prior scores) & School F.E.</i>	0.223	0.186	44.762
<i>Student Fixed Effects</i>	0.105	0.079	50.558
Math Gains with...			
<i>No Controls</i>	0.234	0.234	47.030
<i>Student/Peer Controls</i>	0.232	0.232	47.030
<i>Student/Peer Controls & School F.E.</i>	0.224	0.231	47.030
English Language Arts Levels with...			
<i>No Controls</i>	0.539	0.211	50.335
<i>Student/Peer Controls (incl. prior scores)</i>	0.185	0.162	44.639
<i>Student/Peer Controls (incl. prior scores) & School F.E.</i>	0.175	0.162	44.639
<i>Student Fixed Effects</i>	0.081	0.037	50.335
English Language Arts Gains with...			
<i>No Controls</i>	0.190	0.199	46.242
<i>Student/Peer Controls</i>	0.182	0.195	46.242
<i>Student/Peer Controls & School F.E.</i>	0.175	0.195	46.242

Note: The above estimates are based on the total variance in estimated teacher fixed effects using observations from the pre-experimental data (years 1999-2000 through 2002-03). See the text for discussion of the estimation of the decomposition into teacher by year random effects, student-level error, and "actual" teacher effects. The sample was limited to schools which did not subsequently constitute part of the experimental sample. Any individual students who were in the experiment were dropped from the pre-experimental estimation, to avoid any spurious relationship due to regression to the mean, etc.