Validating Teacher Effect Estimates
Using Changes in Teacher Assignments in Los Angeles[1]

Andrew Bacher-Hicks
Harvard University

Thomas J. Kane[2]
Harvard University and NBER

and

Douglas O. Staiger
Dartmouth College and NBER

Abstract

We evaluate the degree of bias in teacher value-added estimates from Los Angeles using a "teacher switching" design first proposed by Chetty, Friedman, and Rockoff (2014a). We have three main findings. First, as they found in New York City, we find that value-added is an unbiased forecast of teacher impacts on student achievement in Los Angeles, and this result is robust to a range of specification checks. Second, we find that value-added estimates from one school are unbiased forecasts of the teacher's impact on student achievement in a different school, even schools with very different mean test scores. Finally, we find statistically significant differences in average effectiveness of teachers by student race, ethnicity and prior achievement scores that expand gaps in achievement, rather than close them.

Introduction

For nearly half a century, economists have been using panel data to estimate the impacts of teachers and schools on student achievement (e.g., Murnane, 1975; Hanushek, 1971). Because students are sorted to teachers based on student characteristics (such as achievement) which vary over time, analysts have had to rely on the short list of covariates available in education data systems (e.g., prior test scores, student demographics, indicators of free lunch program participation and grade retention) to capture the relevant characteristics used to sort students to teachers and schools (Rothstein, 2010). As such methods migrate from research into policy, the statistical assumptions underlying them are facing greater scrutiny.

In the past few years, various teams of researchers have been testing the predictive validity of teacher and school effect estimates. One widely cited study, Chetty, Friedman and Rockoff (2014a; hereafter CFR), used data from New York City to predict changes in achievement by school and grade based on changes in the average value-added of the teachers assigned to those schools and grades.[3] If a teacher with a large positive value-added estimate leaves a school and joins a new one, one would expect achievement to rise in their new school and to fall in their former school—that is, if the teacher they are replacing is less effective and the estimates were truly attributable to the teacher and not the unmeasured characteristics of the students they left behind.[4] CFR could not reject the hypothesis that the value-added estimates were unbiased predictors of changes in student achievement.

---

[3] CFR (2014a) do not identify the school district. However, during testimony in *Vergara v. California*, the authors subsequently identified the district as New York City.
[4] To avoid measuring achievement with the same data used to predict achievement, CFR (2014a) only used value-added data from outside the two-year window created by any annual change.

In this paper we use the same empirical design as CFR to replicate and extend their analysis. Our analysis is based on seven years of data from Los Angeles Unified School District on over 58,000 teachers and half a million students. We have three main findings.

First, similar to CFR, we find that the teacher-value-added estimates are valid predictors of student achievement when teacher assignments change. We cannot reject the hypothesis that there is no forecast bias, with a confidence interval of ±9%. The heterogeneity in teacher effects is considerably larger in Los Angeles than in New York City. The consequences for math or English achievement of being assigned a top rather than a bottom quartile teacher in Los Angeles are nearly twice as large as CFR found in New York. We further explore the robustness of this result to a number of critiques raised by Rothstein (2014), and subsequently responded to by CFR (2014c). Using data from North Carolina, Rothstein also could not reject the hypothesis of unbiasedness of value-added with a confidence interval of ±5%. However, he found the results sensitive to the handling of missing value-added data. Our results in Los Angeles are not sensitive to various reasonable ways of handling missing value-added data. Rothstein also questioned the quasi-experimental design because he found that changes in the value-added of teachers in a grade were associated with changes in students' prior-year test scores (a "placebo" test). CFR (2014c) argued that these findings were driven by a mechanical relationship, resulting from the fact that prior test scores and the value-added estimates are sometimes estimated with the same data. We replicate the key analyses in CFR (2014c) and Rothstein (2014) and find similar results. In particular, while the main findings are robust to a number of specification changes, the placebo test is not in a way that suggests a mechanical relationship.

Second, we decompose value-added estimates into separate components reflecting a teacher's performance in the same school versus their performance in different schools, and test

for whether the predictive validity of value-added estimates differs for evidence drawn from the same school or from a different school. Recent work by Jackson (2013) has suggested that a given teacher's effectiveness may vary from school to school, depending upon the quality of the match between the teacher and the school. We cannot reject the hypothesis that value-added data is an equally valid predictor whether from the same school, a similar school (as measured by school mean test scores) or a different school.

Finally, unlike CFR, we find statistically significant differences in average effectiveness of the teachers by student race/ethnicity and by prior achievement scores. Teachers differ more in Los Angeles than in the New York, and the allocation of teacher effectiveness in Los Angeles seems to expand gaps in achievement by race/ethnicity and prior achievement, rather than close them. Importantly, unlike CFR, our analysis of the distribution of teacher effects takes teacher experience into account. Many teachers who are effective later in their careers struggle in their early years of teaching, and we find that minority and low-achieving students are more likely to be assigned to novice teachers in Los Angeles. Therefore, it is important to allow for teaching quality to vary by level of experience when estimating the impact of this student-teacher assignment pattern. We find that, in Los Angeles, disparity across students in the combination of experience effects and teacher effects (adjusted for experience) are nearly twice as large as the teacher effects alone.

Testing the Validity of Non-Experimental School and Teacher Effects

In 1986, Robert J. LaLonde compared non-experimental estimates of a training program's impact against the "gold standard" of a randomly assigned comparison group. The

earnings impacts generated using the non-experimental control groups were quite different from those based on the randomized control group.[5] LaLonde's findings have led to a generalized skepticism of non-experimental methods in the study of education and training impacts.

However, it would be inappropriate to generalize the findings to all educational interventions. For instance, the process by which students are sorted to teachers (or schools) and the data available to account for such sorting are quite different from that faced by evaluators when welfare recipients choose a training program. While the reasons underlying a welfare recipient's choice generally remain hidden to a researcher, it is possible that school data systems contain the very data that teachers or principals use to assign students to teachers. Of course, many other unmeasured factors matter for student achievement—such as student motivation or parental engagement. But as long as those factors are also invisible to school principals and teachers when they are making teacher assignment decisions, our inability to control for them would lead to imprecision, not bias.[6]

Unfortunately, given the practical difficulty of randomly assigning students to teachers or to schools, opportunities to replicate LaLonde's comparison of experimental and non-experimental estimates have been rare—until recently. For instance, several recent papers exploit school admission lotteries to compare estimates of the impact of attending a charter school using the lottery-based comparison groups as well as statistical controls to compare charter school and traditional public school students. Abdulkadiroglu et al. (2011) and Angrist, Pathak, and Walters (2013) find similar estimates of the impact of a year in a Boston area charter school whether they use the randomized control group or statistical controls to compare the

---

[5] Dehejia and Wahba (1999) later demonstrate that non-experimental methods perform better when using propensity score methods to choose a more closely matched comparison group.

[6] More problematic would be student- or parent-driven selection of teachers, although the extent of such behavior is hard to measure directly.

performance of students at charter and traditional schools. Deutsch (2012) also finds that the estimated effect of winning an admission lottery in Chicago is similar to that predicted by non-experimental methods. Deming (2014) finds that non-experimental estimates of school impacts are unbiased predictors of lottery-based impacts of individual schools in a public school choice system in Charlotte, North Carolina.

To date, there have been five studies which have tested for bias in individual teacher effect estimates. Four of those—Kane and Staiger (2008), Kane, McCaffrey, Miller and Staiger (2013), Chetty, Friedman and Rockoff (2014) and Rothstein (2014) —estimate value-added for a given teacher in one period and then form empirical Bayes predictions of their students' expected achievement in a second period. The primary distinction between the four studies is the source of the teacher assignments during the second period. In Kane and Staiger (2008), 78 pairs of teachers in Los Angeles working in the same grades and schools are randomly assigned to different rosters of students, which had been drawn up by principals in those schools. The authors cannot reject the hypothesis that the predictions based on teachers' value-added from prior years provide unbiased forecasts of student achievement during the randomized year. However, given the limited sample size, the confidence interval is large, ±35% of the predicted effect.

Kane, McCaffrey, Miller and Staiger (2013) measure teachers' effectiveness using data from 2009-10 and then randomly assign rosters to 1,591 teachers during the 2010-11 school year. The 2009-10 measures include a range of measures, such as value-added, classroom observations and student surveys. The teachers were drawn from six different school districts: New York City (NY), Charlotte Mecklenburg (NC), Hillsborough County (FL), Memphis (TN), Dallas (TX) and Denver (CO). They cannot reject the hypothesis that the predictions based on 2009-10

data are unbiased forecasts of student achievement in 2010-11, following random assignment. The confidence interval for potential bias is ±20%.

Rather than use random assignment, CFR exploit naturally occurring variation in teacher assignments as teachers move from school to school and from grade to grade. Using value-added estimates from other years, they predict *changes* in scores in a given grade and school from t-1 to t based on *changes* in teacher assignments over the same time period. Teacher assignments could change in several different ways: (1) even if the same teachers remain in a school, the proportion of children taught by each teacher could change from t-1 to t; (2) a teacher could exit or enter from a different school; or (3) a teacher could exit or enter from a different grade in the same school. CFR use all three sources of variation to generate their estimates. Each time teacher assignments change from year t to year t-1, CFR have a new opportunity to compare actual and predicted changes in student achievement.

Because they observe many teacher transitions over multiple years, the precision of the estimates in CFR is considerably higher than with either of the previous random assignment studies. Not only can they not reject the hypothesis that the predictions are unbiased, but the confidence interval on their main estimate is much smaller, ±6%.

Rothstein (2014) replicates the CFR findings using data from North Carolina. Using the same methodology, Rothstein cannot reject the hypothesis of unbiasedness with a confidence interval of ±5%.

Glazerman et al. (2013) are the only team so far to use random assignment to validate the predictive power of teacher value-added effects between schools. To do so, they identify a group of teachers with estimated value-added in the top quintile in their state and district. After

7

offering substantial financial incentives, they find a subset of the high value-added teachers willing to move between schools and recruit a larger number of low-income schools willing to hire the high-value-added teachers. After randomly assigning the high value-added teachers to a subset of the volunteer schools, they find that student achievement rose in elementary schools, but not in middle schools. Unfortunately, while their results suggest that teacher value-added estimates have the right sign (at least in elementary schools), they do not investigate whether the magnitude of the impacts are as expected (that is, they could not gauge the magnitude of potential bias).

Methods

Like prior value-added studies, we use a set of control variables generally available in school district administrative data (e.g., prior student achievement, student demographics, average characteristics of students in the class and school average characteristics). However, following CFR, our value-added model differs from prior studies in two key ways:

First, we allow for drift over time as we forecast teacher value-added. Most value-added models assume—either implicitly or explicitly—that a teacher's underlying effectiveness is fixed. Any fluctuations in measured effectiveness are assumed to reflect measurement error, not changes in underlying effectiveness. CFR find evidence to suggest that a teacher's effectiveness has two components: a fixed component and a component representing a true change in effectiveness from one period to the next. As a result, we allow for a similar evolution a teacher's effectiveness in Los Angeles (although, as we note below, there is a greater correlation between teacher effect measures over time in Los Angeles than in New York City.)

Second, we use only within-teacher variation in student, classroom and school-level traits when estimating the influence of those traits on student achievement. Most prior work on value-added models has used a combination of within-teacher and between-teacher variation in these background control variables to adjust for their effects on student achievement. The disadvantage of using both sources of variation is that it becomes impossible to disentangle systematic differences in teacher quality from the influence of the background controls themselves. In other words, when adjusting for student race including between-teacher variation, one is implicitly attributing to student race any possible differences in teacher quality associated with student race. However, by focusing on variation in student traits within teacher and by holding the teacher constant, we preserve the ability to study the relationship between estimated teacher effects and student traits.

Following CFR, we also predict a teacher's impact on students in four steps: First, we estimate the relationship between student test scores and observable characteristics within teachers, using an OLS regression of the form,

$$(1)\ A_{it}^* = \beta X_{it} + \alpha_j + \varepsilon_{ijt}.$$

A_{it}^* represents student i's test score in year t (standardized to have a mean of zero and standard deviation of one), X_{it} includes (1) indicators for gender, race/ethnicity, free and reduced price lunch eligibility, new to school, homelessness, mild or severe special education classification, English language learner classification and prior retention in the current grade; (2) student test scores in both subjects in the prior year (interacted with grade); (3) means of all the demographics and test scores at the school and grade level; and (4) grade-by-year fixed effects. Importantly, α_j is the fixed effect for teacher j.

Second, we calculate the residual student test scores after adjusting for students'

observable characteristics using the following equation:

$$(2)\ A_{it} =\ A_{it}^* - \hat{\beta}X_{it}.$$

The residual student test scores, A_{it}, include the estimated teacher fixed effects, $\hat{\alpha}_j$. We

generate the classroom-level residual, $\bar{A}_{ct,}$ as the mean of the student-level residuals, A_{it}, in each

classroom taught by a given teacher. We then estimate a teacher's effect in a given year, $\bar{A}_{jt}$, as

the precision-weighted average of their classroom-level residuals (with classrooms subscripted

by c below):

$$(3)\ \bar{A}_{jt} = \sum_c w_{ct}\, \bar{A}_{ct},$$

where the precision-weights, w_{ct}, are the inverse of the variance of the estimate of a teacher's

effect from class c.

Third, we estimate drift in value-added flexibly by estimating a different parameter for

each possible time lag s (i.e., school year). In effect, we allow an estimate from five years in the

past to have less predictive value than an estimate from one year in the past. In such cases, the

weight attached to a time lag of length s, γ_s, will be smaller when the absolute value of s is

larger.[7] See CFR (2014a) for details on how the weights, γ_s, are constructed.

Fourth, we put all of these pieces together to estimate different teacher effects, $\hat{\mu}_{jt}$, for

each year t and teacher j, using a "leave-out" approach. Let $\vec{A}_j^{-t}$ indicate the vector of estimates

from years other than year t. Then teacher j's leave-out predicted impact in year t, $\hat{\mu}_{jt}^{-t}$, is simply

---

[7] The teacher fixed effect approach used in prior studies would have granted equal weights to every past year when predicting a teacher's effect next year.

the weighted average of the residuals from years other than year t, with the weights derived from

the drift parameters, γ_s:

$$(4)\ \hat{\mu}_{jt}^{-t} = \gamma' \vec{A}_j^{-t}$$

Data

For this paper, we use information on student demographic characteristics, test scores,

and school and teacher assignments from administrative data provided by Los Angeles Unified

School District. We use data for 7 academic years, from the 2004-05 through the 2010-11.

Before imposing sample restrictions, we observe roughly 1.1 million children and 3.9 million

student-year combinations in grades 3-8. We observe 58 thousand unique teachers and 280

thousand teacher-year combinations in grades 3-8.

Test scores: For those students with a baseline test score in one year, we observe a

follow-up test score for 80% of students in the following spring (this does not include 8th graders

or the last year, spring 2011). We standardize students' scaled test scores to have mean zero and

standard deviation of one by grade and year.

Student Demographics: We use administrative data on a range of other demographic

characteristics for students. These variables include gender, race/ethnicity (Hispanic, white,

black, other or missing), indicators for those ever retained in grade, those eligible for free or

reduced price lunch, those designated as homeless, participating in special education, and

English language learner status.

School and Teacher Assignment Information: We use administrative data indicating

students' grades, schools, and teachers of record (for math and English) in each school year. We

also use the administrative data to derive an indicator for students new to a school and retained in the current grade.

Sample Restrictions: To construct an analysis sample, we use a series of sample restrictions that closely mirror other value-added work. First, we include students in grades 3-8 who could be linked to a math or English teacher. Second, we exclude the 2% of observations in classrooms where more than 50% of the students were identified as special education students. Third, we remove classrooms with extraordinarily large (more than 45) or extraordinarily small (less than 5) enrolled students, which excludes 1% of students with valid scores.

After these sample restrictions, we have 3 million observations with information on teacher assignments, test score gains and demographics. We combine all of these data elements together into one dataset with one row per student per subject (math or English) per school year. Each row contains the student test score (both current and prior year), student demographic information, and school and teacher assignment information.

Sample Statistics from Combined Data: We report sample statistics for relevant data and student characteristics in Table 1. The first two rows present information on the number of unique students and classrooms. We observe 591,803 unique students, with an average of 5.08 subject-school years. We observe 141,853 unique classrooms, with an average class size of 24.37 students.

We standardize student test scores on the full sample of students (before any sample restrictions), including special education students. As a result, the analysis sample is slightly higher achieving and slightly less diverse than the population. The average standardized test score is slightly larger than zero, .13, and the standard deviation is slightly smaller than one,

.983.   A high percentage of students (78%) in Los Angeles are eligible for free and reduced-priced lunch. There is also a high percentage of Hispanic students (75%) and high percentage of students who have limited English proficiency (28%).

Heterogeneity and Drift in Teacher Effects

Our estimate of the heterogeneity in teacher effects is considerably larger in Los Angeles than CFR's estimate for New York City. CFR find that a standard deviation in teacher impacts is equivalent to .124 and .163 student-level standard deviations in achievement in elementary school English and math respectively, and .079 and .134 in middle school (CFR 2014a, Table 2, Panel B). The comparable estimates in Los Angeles are .189 and .288 in elementary English and math respectively, and .097 and .206 in middle school English and math. There are many reasons that variance in teacher effectiveness might be higher in Los Angeles than in other urban district. In particular, Los Angeles has traditionally had a more centralized hiring process, which gives principals less authority in selecting new hires and has a shorter probationary period before teachers get tenure. It could also be that their testing outcomes are more sensitive to teacher influences.

In Figures 1 and 2, we present the distribution of predicted teacher effects for elementary and middle school teachers respectively. The distribution of predicted teacher effects is somewhat narrower than the underlying differences in teacher effects reported in the paragraph above, because the prediction method "shrinks" each teacher's estimate toward zero. Nevertheless, the implied difference in effectiveness between the most effective and least effective teachers is quite large, especially in mathematics. For instance, a teacher at the 75th percentile or above is predicted to raise achievement in his or her class by .3 student-level

standard deviations, relative to the average classroom in elementary math. Conversely, a teacher at the 25th percentile is predicted to reduce student achievement by a similar amount.

CFR (2014a) report comparable figures to Figures 1 and 2 in their Appendix Figure 1. The standard deviation of predicted teacher impacts in the district they studied is .080 and .116 in elementary English and math respectively, and .042 and .092 in middle school. Assuming a normal distribution, the predicted impact of being assigned a top quartile teacher in elementary math is .145 standard deviations, roughly half as large as the comparable estimate in Los Angeles.

In Figure 3, we report the correlations between the weighted mean achievement residuals by teacher and year, $\bar{A}_{jt}$, at various time lags. There is considerably less drift in the teacher effect estimates in Los Angeles relative to New York City. For example, in elementary math, we find a correlation in teachers' weighted-mean residuals one year apart of .66. The comparable figure in CFR is .43. The higher correlations over time likely reflect the greater heterogeneity in underlying teacher effects in Los Angeles.

Changes in Student Achievement following Changes in Teacher Assignments

Following CFR, we test the predictive validity of the value-added estimates by studying the changes in achievement in each school and grade corresponding with changes in teacher assignments. More precisely, we predict the change in average student achievement given changes in the weighted average of teacher value-added in that school and grade. The average predicted value-added, Q_{sgst}, is simply the weighted average of the predicted teacher effects,

$\hat{\mu}_{jt}^{-\{t,t-1\}}$, for the teachers assigned to that grade and subject, weighted by their enrollments.[8] The change from year t-1 to t is ΔQ_{sgst}. We calculate the change in average raw test scores at each school-grade-subject-year cell from year t-1 to t, ΔA^*_{sgst}, and then estimate the following equation:

$$(5) \quad \Delta A^*_{sgst} = \beta_0 + \beta_1 \Delta Q_{sgst} + \varepsilon_{sgst}$$

In Table 2, column 1 is the preferred specification from CFR (2014a). They report a parameter estimate of .974 and a standard error of .033, which implies a forecast bias of -2.6% (100-97.4) and a confidence interval of ±6.5%. Our estimate in column 1 is quite similar, 1.030, and implies a forecast bias of 3.3% (103.3-100). The confidence interval around the estimate is ±8.6% (±1.96*.044).

Figure 4 presents the graphical version of the results in column 1. First, we sort each school-grade-subject-year cell into one of 20 groups, based on the magnitude of the predicted change in value-added, ΔQ_{sgst}. Then we calculate the average change in actual scores in each of the 20 groups, ΔA^*_{sgst}. In Figure 4, we present the scatter plot of the means of predicted change in scores and actual change in scores for all 20 groups. Two facts are evident. First, the changes in actual scores match the changes in predicted scores throughout the distribution. Second, especially at the tails, the magnitude of the change is quite large. Figure 4 reports changes in average scores for whole grade levels within a school. In 10% of all school-grade-subject cells, we would have predicted changes in scores of ±.15 standard deviations based simply on changes in the teacher assignments in those school-grade-subject cells (5% of cells predicted to have an

---

[8] Since this section focuses on *changes* from t-1 to t, we only used the teacher effect estimates for the years outside the two-year window, t to t-1, to form the predictions.

increase of .15 and 5% with a decrease of .15 standard deviations). The results suggest that the average change in actual achievement roughly corresponded with those predictions.

Columns 2 and 3 of Table 2 report results separately for middle school grades (6-8) and elementary grades (4 and 5) respectively. The coefficients for middle school and elementary school, 1.122 and .996, are not statistically distinguishable from one.

The last two columns of Table 2 report various robustness checks. One concern is that teacher turnover may coincide with other changes in a school. As a result, instead of imposing an assumption that the year effects are common across schools, column 4 allows for different year effects by school. In effect, these estimates are only relying on *differential* changes in scores and predicted value-added by grade and subject within a school, since the mean change in scores and predicted value-added that is shared across multiple grades and subjects is being subtracted out. The coefficient is .963 with confidence interval of ±9%. Column 5 adds controls for changes in the predicted mean value-added of teachers in the school-grade-subject in the prior and subsequent years. The coefficient is .942 with a confidence interval of ±11%. In all of these specifications, the confidence interval contains one and does not include zero.[9]

Teachers with Missing Value-Added

Throughout most of their analysis, CFR exclude from consideration classrooms taught by teachers with no value-added estimate outside of the two-year window. In Table 2, we have applied the same restriction. However, as a robustness check, CFR include teachers with

---

[9] In addition to controls for school-by-year fixed effects and lead and lag changes in teacher value-added, CFR include a specification that controls for the change in scores for the same subject and other subject in the prior year. We also replicated this finding, but do not report the changes in Table 2, since it is not appropriate to include this control, as we discuss in the section on the use of lagged scores below. For comparative purposes, when estimating a model with school-by-year fixed effects, controls for lead and lagged changes in teacher value-added, and lagged score controls, we estimate a coefficient on changes in mean across cohort value-added of .87, with a confidence interval of ±7%.

missing value-added data, imputing their value-added to be zero (i.e., attributing to the missing teachers the mean teacher effectiveness). When doing this, the coefficient on predicted achievement falls to .877, an estimate which is statistically different from one (CFR 2014a, Table 5, column 2). The authors interpret the decline as being attributable to measurement error.

In Table 3, we report the preferred specification from CFR in column 1 and then apply several alternative approaches to imputing value-added for those with missing values. First, we assign the whole-sample mean effectiveness, 0, to any teacher with missing value-added. As reported in column 2, we find an estimate of .993, with a confidence interval of ±10%. In other words, the estimates in Los Angeles are less sensitive to the assumption of average value-added than in the district studied by CFR. For column 3, we re-estimate equation (1) including controls for teacher experience, with indicators for each single year of experience from one through nine years and one additional indicator for teachers with 10 or more years of experience. Therefore, in addition to $\hat{\mu}_{jt}^{-\{t,t-1\}}$, we can use teaching experience to impute value-added for those with missing $\hat{\mu}_{jt}^{-\{t,t-1\}}$. The coefficient is .996 with a confidence interval of ±9%. Next, we exploit the fact that many teachers with missing value-added outside the two-year window had value-added estimates during the window (for example, early career teachers who leave before their third year would have value-added in their first two years but would necessarily have missing two-year leave-out value-added for all years). The grand mean value-added for these teachers is -.049 during the two-year window. Therefore, in column 4 of Table 3, we use -.049 to impute value-added for missing teachers. The coefficient is essentially unchanged at .998 with a confidence interval of ±10%. Finally, in column 5 we perform the simple exercise of restricting the sample to only include cells where no teachers are missing two-year leave-out value-added estimates. Again, the coefficient remains substantially unchanged at .973 with a

confidence interval of ±9%. Based on these findings, we conclude that the treatment of teachers with missing value-added has little effect on the estimates in Los Angeles.

Additional Robustness Checks

Table 4 presents two more robustness checks. The first two columns add changes in predicted effectiveness of teachers in the <u>other subject</u> in a grade level and school. Changes in predicted effectiveness in other subjects capture underlying changes in the quality of teaching in the school, such as might occur with changes in school leadership. Column 1 reports the results for grades 6-8, while column 2 reports results for grades 4 and 5. In the grades 6-8, where teachers generally specialize by subject, those teaching other subjects are literally different people. A positive coefficient implies that there is some evidence of "spillover". For instance, in middle school, when the quality of teaching improves in one subject, achievement does seem to improve in the other subject as well, by .282 standard deviations with a confidence interval of ±20% (which excludes zero). However, the coefficient on "own subject" remains at 1.078 with a confidence interval which includes one (implying that the changes in effectiveness in the other subject are not highly correlated with changes in effectiveness within a given subject). In elementary school, the coefficient is .160, but more precisely estimated with a confidence interval of ±5%. This is not surprising, since elementary teachers typically teach multiple subjects to the same students. Also, the coefficient on "own subject" falls significantly below 1 to .904. However, this result reflects the fact that our predictions of teacher value-added in each subject only use information from the teacher's performance in the same subject. In a more complete model of elementary teachers, the prediction of teacher value-added in each subject would depend on the teacher's value-added in both subjects (Lefgren and Sims, 2012). Thus,

when we include other subject value-added in elementary, other subject value-added receives some weight while own subject value-added receives less weight.

The change in predicted teacher effectiveness can arise from a number of different types of changes—changes in the proportion of students taught by each teacher in a grade and subject, teachers switching from one grade to another, or teachers exiting or entering a school. The key assumption in the CFR methodology is that teachers are not sorting to students in the same way from year to year. This seems safest to assume when a teacher leaves or enters a school, since the new teachers will typically be unfamiliar with the principal and students. As a result, for the instrumental variable estimate in column 3, we instrument for ΔQ_{sgst} by multiplying the fraction of students in the prior year's school, grade, subject, year cell taught by teachers who leave the school by the mean effectiveness estimates of these leavers. Therefore, the estimates in column 3 of Table 4 are focusing on the variation in teacher effectiveness driven by teacher exit. Still, the coefficient is not statistically different from one, .972, with a confidence interval of ±16%.

The Lagged Score "Placebo" Test

There is no control for the change in student baseline scores in equation (5). Like CFR, we are effectively assuming that the change in predicted value-added is exogenous to any change in baseline achievement. In a recent paper, Rothstein (2014) reports a statistically significant relationship between change in teacher value-added and changes in baseline achievement as *prima facie* evidence of bias in the CFR method. When we replicate his analyses, we similarly find that the predicted change has a coefficient of .268 when lagged scores are the dependent variable. However, rather than invalidating the CFR methodology, CFR (2014c) argue that the lagged score test merely demonstrates the hazards of using the same data to estimate the

19

dependent and independent variables. There is a mechanical relationship between the two, which enters through two routes. First, because teachers frequently switch grades in a school from one year to the next, the value-added predictions will be based on the some of the very same data included in the baseline scores. CFR's two-year leave-out window is designed to resolve this problem when ΔA^*_{sgst} is the dependent variable. Rothstein reintroduces the problem when he uses ΔA^*_{sgst-1} as the dependent variable. If a school sees a large improvement in the predicted value-added of teachers in grade g, some of the new teachers will have just taught grade g-1 in the previous year. Second, Kane and Staiger (2002) document the existence of school by subject-year random effects, which could also produce a relationship between ΔQ_{sgst} and ΔA^*_{sgst-1}. If these shocks are serially correlated, such a relationship could persist even with a three-year leave-out window.

Accordingly, in Table 5, we replicate a number of specifications from both Rothstein and CFR to explore the relationships between changes in value-added and lagged scores. The table reports the coefficients on the change in average value-added, ΔQ_{sgst}, with a range of different specifications. For the specifications in the top row, the dependent variable is change in end-of-year achievement, ΔA^*_{sgst}; in the bottom row, the dependent variable is the change in lagged score (or baseline score), ΔA^*_{sgst-1}. Across all of our specifications, we find that the coefficient is stable and indistinguishable from one when change in end-of-year achievement, ΔA^*_{sgst}, is the dependent variable. However, we find that the coefficient is sensitive to the model specification when change in lagged achievement, ΔA^*_{sgst-1}, is the dependent variable.

Column 1 replicates CFR's preferred specification. When the change in end of year scores is the dependent variable, the coefficient is indistinguishable from one. When the change

in lagged scores is the dependent variable, the coefficient is .268, with a confidence interval of ±8%. In columns 2 and 3, we use a three-year leave-out window (leaving out the two prior years and the current year), which was the solution Rothstein (2014) proposed to solve the mechanical relationship introduced in his initial placebo test. In column 2, the coefficient is .246 with a confidence interval of ±9%. In column 3, we include fixed effects by school, year and subject, and the coefficient is 0.212, with a confidence interval of ±10%. In columns 4 through 6, we present analyses suggested by CFR (2014c) to solve the mechanical relationship with lagged scores. In column 4, we instrument for the change in value-added excluding those teachers who taught in the previous grade in the previous year. The coefficient when change in lagged scores is the dependent variable remains statistically significant, but has fallen to .178. In column 5, we add fixed effects by school, year and subject. The coefficient when change in lagged scores is the dependent variable again falls to .105, although it remains statistically significant. In column 6, we instrument for the change in lagged score excluding teachers who ever switched grades within a school. Under this final specification, the coefficient on lagged score is .049 with and a confidence interval of ±11% and is not statistically significant.

In sum, throughout Table 5, all of the coefficients on end of year score are statistically significant and none are distinguishable from one, yet the coefficients on lagged score are much more sensitive to the model specification. This leads us to conclude that any potential forecast bias suggested by this test is minimal and likely operating through a combination of teacher switching and shared school-by-year-by-subject effects. When we take steps to adjust for these sources of a mechanical relationship, the coefficient on predicted value-added goes to zero when predicting baseline scores, but remains equal to one in predicting end of year scores.

Are Teacher Value-Added Estimates Context-Specific?

Even if value-added estimates are unbiased predictors within a given school environment, the same teacher could be more or less effective in a different school. Using data from North Carolina, Jackson (2013) estimates that teacher-school match effects account for roughly one-third of the variance in teacher effects. To test the possibility that teacher effects vary by context, we first divide the data available for each teacher's value-added into value-added estimates using observations only from the same school and those from all available schools. Suppose $\hat{\mu}_{j,same}^{-\{t,t-1\}}$ represents the teacher effect estimate for teacher j using only data from the same school and $\hat{\mu}_{j,all}^{-\{t,t-1\}}$ the estimates from the full dataset. Because data from the same school are just a subset of the data from all schools, the law of iterated expectations implies that $\hat{\mu}_{j,same}^{-\{t,t-1\}}$ and $\hat{\mu}_{j,all}^{-\{t,t-1\}} - \hat{\mu}_{j,same}^{-\{t,t-1\}}$ are orthogonal, and the latter term represents the component of $\hat{\mu}_{j,all}^{-\{t,t-1\}}$ that reflects the additional information from teacher performance in others schools. Therefore, in Table 6, we include both $\hat{\mu}_{j,same}^{-\{t,t-1\}}$ and $\hat{\mu}_{j,all}^{-\{t,t-1\}} - \hat{\mu}_{j,same}^{-\{t,t-1\}}$. Consistent with Jackson's findings regarding teacher-school matches, the point estimate on value-added estimate of the same school, 1.054, is larger than the coefficient on the difference between that from all schools and similar schools, .817. However, we cannot reject the hypothesis that both coefficients were equal to one (p value=.163).

In column 2 of Table 6, we divide the data for each teacher and year into three sequentially nested groups: data from the same school, data from the same or similar schools (with mean scores within .1 standard deviations) and data from all schools. We use such data to create three orthogonal variables: $\hat{\mu}_{j,same}^{-\{t,t-1\}}$, $\hat{\mu}_{j,similar}^{-\{t,t-1\}} - \hat{\mu}_{j,same}^{-\{t,t-1\}}$ and $\hat{\mu}_{j,all}^{-\{t,t-1\}} - \hat{\mu}_{j,similar}^{-\{t,t-1\}}$. Again, the coefficient on the teacher effect estimates from the same school, 1.054, is larger than the coefficient on latter two differences, .760 and .838. Nevertheless, we cannot reject the

hypothesis that the coefficients were all equal to one (p-value=.275). In other words, we cannot reject the hypothesis that a teacher's value-added estimate from a different school or from a school with considerably higher or lower mean test scores were equally predictive of their students' achievement.

The Distribution of Teaching Effectiveness (Including Teaching Experience)

A central question in current policy debate is the degree to which different groups of students have access to the same quality teaching. For instance, in *Vergara v. California,* the plaintiffs argued that the least effective teachers were disproportionately assigned to low-income, minority and lower achieving students. When testing for differences in mean teacher effectiveness by student characteristics, the empirical challenge is to disentangle systematic differences in teacher quality from the direct effects of those same student characteristics. For instance, if one were to estimate equation (1) without teacher fixed effects, and included class-level or school-level mean baseline achievement among the covariate controls, X, there would be no relationship between teacher value-added and the covariates in X. However, this would be true by construction, since value-added is calculated as the residual from equation (1). An important strength of the CFR methodology is that by including teacher fixed effects in equation (1), they preserve the possibility that teacher effects could be correlated with the covariate controls, X.

Following CFR, Table 7 uses the teacher effect estimates, $\hat{\mu}_{jt}$, as the dependent variable and estimates the relationship between teacher effectiveness and observable student characteristics both at the student-level and the school-level. Column 1 presents estimates of the relationship between teacher effect estimates, $\hat{\mu}_{jt}$, and student-level prior-year test scores, $A_{i,t-1}^*$.

The point estimate of .024 is statistically significant, and implies that a one-standard deviation increase in prior achievement is associated with being assigned a teacher with .024 higher predicted effectiveness. In other words, rather than being used to narrow achievement gaps, teacher assignments in Los Angeles exacerbate prior achievement differences, with weaker students being assigned weaker teachers. The point estimate is roughly three times as large as the point estimate observed by CFR.

In column 2, we include fixed effects by school when estimating the relationship between teacher effect estimates, $\hat{\mu}_{jt}$, and student-level prior-year test scores, $A^*_{i,t-1}$. With school fixed effects included, the estimates are based on differences within school. The point estimate of .013 is statistically significant, but smaller than the estimate without school fixed effects. The results presented in columns 1 and 2 imply that students with higher prior-year test scores are a) in schools with higher value-added teachers and b) that within schools, students with relatively higher prior-year test scores (than other students in that same school) are placed with teachers with relatively higher value-added (than other teachers in that same school).

In column 3, we present the analogous estimates by student race/ethnicity. Relative to white students, African-American, Asian, and Hispanic students in Los Angeles are assigned less effective teachers, on average. African-American students are assigned teachers with average value-added .030 student-level standard deviations below the average of teachers to whom white students are assigned. In other words, the average African-American student in Los Angeles is losing .030 standard deviations in achievement each year relative to white students with similar prior achievement, because of the lower effectiveness of the teachers to which they are assigned. Latino students, who comprise 75% of students in Los Angeles, are losing .043 standard

deviations per year relative to similar white students in Los Angeles, because of the teachers they are assigned.

In column 4, we present the results by race/ethnicity after adding fixed effects by school. The estimates are statistically significant and negative for African-American and Latino students, but they are much smaller, -.010 rather than -.030 and -.043 respectively.[10] In other words, much of the difference in teacher quality by race/ethnicity is due to the mal-distribution of teacher effectiveness between schools, although there is still evidence that African-American and Latino students are assigned less effective teachers within the same schools. The results also imply that the difference between white and Asian students is entirely due to between-school differences. Within schools, the white-Asian difference is not statistically significant.

Column 5 illustrates a similar point by regressing teacher value-added against the fraction of students in each school in various racial/ethnic groups. Schools with more African-American and Latino students have lower teacher quality, on average.

One weakness in the CFR methodology is that in estimating, $\hat{\mu}_{jt}$, they take no account of teaching experience. However, in many school districts, teachers start their careers teaching more disadvantaged students and, as they gain experience, move to teaching higher income and higher-achieving students (Boyd, Loeb and Wyckoff, 2008; Kalgorides, Loeb and Beteille, 2013; Jackson, 2013). Value-added estimates typically rise sharply during teachers' first several years of teaching and then flatten out afterward. Failing to account for teacher underperformance during the early years of teaching may understate the differences in teacher quality for more and less advantaged students.

---

[10] As CFR note, these estimates *understate* the differences in true value-added across groups since the dependent variable is a 'shrunken' estimate of true teacher value-added.

To investigate this possibility, we re-estimate equation (1) including 10 indicators of a teacher's number of years of experience (we used an indicator variable for each of the first nine years of experience and one indicator for all teachers with 10 or more years of experience). Instead of using $\hat{\mu}_{jt}$ as the dependent variable, the top panel of Table 8 uses the teachers' experience multiplied by the relevant experience effect as the dependent variable. As reported in column 1, there is a .017 standard deviation difference in teacher effectiveness per standard deviation in student baseline achievement based simply on teaching experience. Analogously, African-American and Latino students are losing .039 and .018 standard deviations respectively relative to white students based simply on the differential in the average experience of their teachers. Most of the experience effects seem to be operating at the school level, as only the difference associated with the baseline achievement, .004, is statistically significant after including school effects in columns 2 and 4.

In the bottom panel of Table 8, we use the sum of the adjusted teacher effects, $\hat{\mu}_{jt}$, (which have been re-estimated to adjust for the teacher experience effects) and the experience effects as the dependent variable. While $\hat{\mu}_{jt}$ may be useful for summarizing the differences in "teacher" effectiveness, the sum of $\hat{\mu}_{jt}$ and the experience effects is a better measure of the difference in "teaching effectiveness"—acknowledging the fact that the average teacher improves during their initial years of teaching. The combined effects are substantially larger than in Table 7. Rather than a .024 difference in teacher effectiveness per standard deviation in baseline performance, the difference is .042 standard deviations in teaching effectiveness, once experience effects are included. The deficit in teaching effectiveness for African American and Latino students relative to white students is .069 and .063 standard deviations respectively.

Conclusion

There is now substantial evidence that non-experimental teacher effect measures (often called "value-added" measures) capture important information about the causal effects of teachers on student achievement. Since 2008, three studies using random assignment in different sites have confirmed the validity of teacher-level value-added estimates (Kane and Staiger, 2008; Kane, McCaffrey, Miller and Staiger, 2013; Glazerman, Protik, Teh, Bruch, Max, and Warner, 2013). In addition, the CFR methodology has produced little evidence of bias in three sites so far: New York City, Los Angeles and North Carolina. Rarely in social science have we seen such a large number of replications in such a short period of time. Even more rare is the high degree of convergence in the findings.

Despite the lack of evidence of prediction bias, questions linger in three areas:

First, we have much to learn about the role of school context and "match quality" in teacher effect estimates. Although we cannot rule out the hypothesis that teacher effect estimates derived from a teacher's experience in another school were equally valid predictors as the same-school estimates, we have too little power to rule out Jackson's (2013) findings on the importance of match quality as a component of teacher effectiveness.

Second, the field must narrow the range of statistical specifications which practitioners should expect to yield valid predictions. For instance, it has become conventional among economists to condition on classroom and school mean characteristics when estimating value-added models. However, very few district or state systems currently include controls for peer effects when generating value-added estimates for their employment and pay decisions. Unfortunately, we have little statistical power using the CFR method to test the importance of

peer controls simply because of the limited variation in peer control variables when aggregated at the school-by-year level. We need more studies specifically designed to test for the importance of peer controls and other specification decisions.

Finally, although none of the validity studies so far have produced evidence of bias, we know very little about how the validity of the value-added estimates may change when they are put to high stakes use. All of the available studies have relied primarily on data drawn from periods when there were no stakes attached to the teacher value-added measures. In the coming years, it will be important to track whether or not the measures maintain their predictive validity as they are used for tenure decisions, teacher evaluations and merit pay.

Bibliography:

Abdulkadiroglu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane and Parag
      Pathak. (2011) "Accountability and Flexibility in Public Schools: Evidence from
      Boston's Charters and Pilots." *Quarterly Journal of Economics* 126(2): 699–748.

Angrist, Joshua D., Parag Pathak and Christopher R. Walters. (2013) "Explaining Charter
      School Effectiveness." *American Economic Journal: Applied Economics* 5(4): 1–27.

Boyd, Donald, Hamilton Lankford, Susanna Loeb and James Wyckoff. (2008) "The narrowing
      gap in New York city teacher qualifications and its implications for student achievement
      in high-poverty schools." *Journal of Policy Analysis and Management* 27(4): 793-818

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. (2014a) "Measuring the Impacts of
      Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic
      Review* 104(9): 2593-2632.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. (2014b) "Prior Test Scores Do Not Provide
      Valid Placebo Tests of Teacher Switching Research Designs." Unpublished research
      note. http://obs.rc.fas.harvard.edu/chetty/va_prior_score.pdf.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. (2014c) "Response to Rothstein (2014)
      'Revisiting the Impacts of Teachers.'" Unpublished research note.
      http://obs.rc.fas.harvard.edu/chetty/Rothstein_response.pdf

Dehejia, Rajeev H. and Sadek Wahba. (1999) "Causal Effects in Nonexperimental Studies:
      Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical
      Association* 94(448): 1053-1062.

Deming, David. (2014) "Using School Choice Lotteries to Test Measures of School
      Effectiveness." *American Economic Review: Papers & Proceedings* 104(5): 406–411.

Deutsch, Jonah. (2012) "Using School Lotteries to Evaluate the Value-Added Model."
      University of Chicago Working Paper.

Glazerman, Steve, Ali Protik, Bing-ru The, Julie Bruch, Jeffrey Max and Elizabeth Warner.
      (2013). "Transfer Incentives for High-Performing Teachers: Final Results from a
      Multisite Experiment" (NCEE 2014-4003). Washington, DC: National Center for
      Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S.
      Department of Education.

Hanushek, Eric A. (1971) "Teacher Characteristics and Gains in Student Achievement:
      Estimation Using Micro-Data." *American Economic Review* 61(2): 280-288.

Jackson, Kirabo. (2013) "Match Quality, Worker Productivity, and Worker Mobility: Direct
      Evidence From Teachers." *Review of Economics and Statistics* 95(4): 1096-1116

Kalogrides, Demetra, Susanna Loeb and Tara Beteille. (2013) "Systematic sorting: Teacher characteristics and class assignments." *Sociology of Education* 86(2): 103-123.

Kane, Thomas J. (2004) "The impact of after-school programs: Interpreting the results of four recent evaluations." Working paper. New York: William T. Grant Foundation.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. (2013) "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." Seattle, WA: Bill & Melinda Gates Foundation.

Kane, Thomas J. and Douglas O. Staiger (2002) "The Promise and Pitfalls of Using Imprecise School Accountability Measures" *Journal of Economic Perspectives* 16(4): 91-114.

Kane, Thomas J., and Douglas O. Staiger. (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper 14607.

LaLonde, Robert J. (1986) "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76(4): 604-20.

Lefgren, L, Sims, D. (2012) "Using Subject Test Scores Efficiently to Predict Teacher Value-Added." *Education Evaluation and Policy Analysis* 34(1): 109-121.

Murnane, Richard J. (1975) The Impact of School Resources on the Learning of Inner City Children (Cambridge, Mass.: Ballinger Publishing)

Rothstein, Jesse. (2010) "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement" *The Quarterly Journal of Economics* 125 (1): 175-214.

Rothstein, Jesse. (2014) "Revisiting the Impacts of Teachers." Unpublished working paper. http://eml.berkeley.edu/~jrothst/workingpapers/rothstein_cfr.pdf

Table 1. Summary Statistics

	Mean	SD	Observations
Dataset Characteristics:			
Number of subject-school years per student	5.08	2.91	591,803
Class size (not student-weighted)	24.37	6.25	141,853
Student Characteristics:			
Test Score (student-level standard deviation units)	0.13	0.98	3,008,965
Male	49.55%	0.50	3,009,024
African-American	9.01%	0.29	3,009,024
Asian	6.79%	0.25	3,009,024
Hispanic	75.11%	0.43	3,009,024
White	9.09%	0.29	3,009,024
Repeating Grade	0.07%	0.03	3,009,024
Free or Reduced Price Lunch Eligible	77.63%	0.42	3,009,024
Homeless	1.08%	0.10	3,009,024
Mild SPED	4.82%	0.21	3,009,024
Severe SPED	1.21%	0.11	3,009,024
ELL - Reclassified to Fluent English Proficient	29.93%	0.46	3,009,024
ELL - Limited English Proficient	28.32%	0.45	3,009,024
ELL - Initially Fluent English Proficient	12.17%	0.33	3,009,024

Notes: This sample of students and teachers is limited to those with the requisite information for estimating value-added (e.g., students must have prior-year test scores). For more discussion of these sample restrictions, see the sample restrictions section of the paper.

Table 2. Quasi-Experimental Estimates of Forecast Bias

	Same Subject				
	b / (se)	b / (se)	b / (se)	b / (se)	b / (se)
Changes in mean teacher VA across cohorts	1.030	1.122	0.996	0.963	0.942
	(0.044)	(0.131)	(0.037)	(0.048)	(0.055)
Year Fixed Effects	Yes	Yes	Yes		
School-by-Year Fixed Effects				Yes	Yes
Lagged and Forward Teacher VA					Yes
Grades	4 to 8	6 to 8	4 and 5	4 to 8	4 to 8
N	14,186	3,434	10,752	14,186	9,170

Notes: Each column reports coefficients and standard errors (in parentheses) from an OLS regression, where the dependent variable is the change in mean test scores within a school-grade-subject cell from the prior year to the current year. Standard errors are clustered at the school-cohort level. The regressions are estimated in a dataset aggregated to school-grade-subject-year cells and weighted by the number of students in the school-grade-subject-year cell. Classrooms in which two-year leave-out value-added estimates cannot be constructed are excluded.

Table 3. Quasi-Experimental Estimates of Forecast Bias Robustness Check: Sensitivity to Missingness

	Missing VA Excluded (Main Model) b/ (se)	Missing VA set to 0 b/ (se)	Missing VA Imputed by Teaching Experience b/ (se)	Missing VA set to -.049 (average residual of teachers with missing leave-out VA) b/ (se)	Only Cells with No Teachers Missing VA b/ (se)
Changes in mean teacher VA across cohorts	1.030	0.993	0.996	0.998	0.973
	(0.044)	(0.049)	(0.048)	(0.049)	(0.048)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
N	14,186	14,292	14,292	14,292	8,974

Notes: Each column reports coefficients and standard errors (in parentheses) from an OLS regression, where the dependent variable is the change in mean test scores within a school-grade-subject cell from the prior year to the current year. Standard errors are clustered at the school-cohort level. The regressions are estimated in a dataset aggregated to school-grade-subject-year cells and weighted by the number of students in the school-grade-subject-year cell. Column 1 repeats the same specification reported in column 1 of Table 2, which excludes teachers for whom two-year leave-out value-added estimates cannot be constructed. Column 2 includes these teachers, by setting their VA to 0, the full-sample VA mean. Column 3 includes these teachers by imputing VA based on experience. Column 4 includes these teachers by setting their VA to -0.49, which is the average VA of teachers for whom two-year leave-out VA cannot be calculated. Column 5 includes only cells where no teachers are missing two-year leave-out value-added estimates.

Table 4. Additional Robustness Checks: Predicted Effectiveness in Other Subjects and Using Teacher Exit as an Instrument

	Other Subject		Teacher Exit Only (IV)
	b / (se)	b / (se)	b / (se)
Changes in mean teacher VA across cohorts	1.078	0.904	0.972
	(0.126)	(0.031)	(0.082)
Change in mean teacher other subject VA across cohorts	0.282	0.160	
	(0.100)	(0.026)	
Year Fixed Effects	Yes	Yes	Yes
Grades	6 to 8	4 and 5	4 to 8
N	3,394	10,752	14,186

Notes: Each column reports coefficients and standard errors (in parentheses) from an OLS regression (columns 1 and 2) or a 2SLS regression (column 3), where the dependent variable is the change in mean test scores within a school-grade-subject cell from the prior year to the current year. Standard errors are clustered at the school-cohort level. The regressions are estimated in a dataset aggregated to school-grade-subject-year cells and weighted by the number of students in the school-grade-subject-year cell. Classrooms in which two-year leave-out value-added estimates cannot be constructed are excluded. Columns 1 and 2 restrict the sample to middle school and elementary school, respectively and control for other subject changes in mean teacher VA across cohorts in addition to same subject. Column 3 reports estimates from a 2SLS regression, instrumenting for changes in mean teacher VA with the fraction of students in the prior cohort taught by teachers who leave the school multiplied by the mean VA among those teachers.

Table 5. The Lagged Score Placebo Test

Dependent Variable:	Baseline Model (Two-Year Leave-Out)	Baseline Model with Three-Year Leave-Out		No Followers (IV)		No Within-School Movers (IV)
	b / (se)	b / (se)	b / (se)	b / (se)	b / (se)	b / (se)
Change in Current Score	1.030	0.991	0.964	0.995	0.950	0.963
	(0.044)	(0.048)	(0.060)	(0.049)	(0.042)	(0.056)
Change in Lagged Score	0.268	0.246	0.212	0.178	0.105	0.049
	(0.039)	(0.043)	(0.053)	(0.044)	(0.041)	(0.055)
Year Fixed Effects	Yes	Yes		Yes		
School x Year x Subject Fixed Effects			Yes		Yes	Yes
N	14,186	14,186	14,186	14,186	14,186	14,186

Notes: Each column reports coefficients and standard errors (in parentheses) from an OLS (columns 1-3) or a 2SLS (columns 4-6) regression, where the dependent variable is the change in mean test scores within a school-grade-subject cell from the prior year to the current year. Standard errors are clustered at the school-cohort level. The regressions are estimated in a dataset aggregated to school-grade-subject cells and weighted by the number of students in the school-grade-subject-year cell. Classrooms in which two-year leave-out value-added estimates cannot be constructed are excluded. Column 1 repeats the same specification reported in column 1 of Table 2, which includes all teachers for whom two-year leave-out value-added estimates exist. Columns 2 and 3 also repeat the same specification reported in column 1 of Table 2, but exclude an additional prior year of data (i.e., t-2, t-1, and t are excluded, instead of just t-1 and t). Columns 4 and 5 report estimates from a 2SLS regression, instrumenting for changes in mean teacher VA with the changes in mean VA excluding teachers who switch from the previous grade to current grade (i.e., they 'follow' the students). Column 6 reports estimates from a 2SLS regression, instrumenting for changes in mean teacher VA with the changes in mean VA excluding teachers who switch across grades within a school.

Table 6. Predicted Effectiveness from Using Estimates from Same, Similar, and Different Schools

	b / (se)	b / (se)
VA Same School	1.054	1.053
	(0.046)	(0.046)
(VA All Schools) - (VA Same School)	0.817	
	(0.125)	
(VA Similar Schools) - (VA Same School)		0.760
		(0.318)
(VA All Schools) - (VA Similar Schools)		0.838
		(0.124)
Joint Test That All Coefficients Equal 1 (p-value)	(0.163)	(0.275)
Year Fixed Effects	Yes	Yes
N	14,182	14,182

Notes: Each column reports coefficients and standard errors (in parentheses) from an OLS regression, where the dependent variable is the change in mean test scores within a school-grade-subject cell from the prior year to the current year. Standard errors are clustered at the school-cohort level. The regressions are estimated in a dataset aggregated to school-grade-subject-year cells and weighted by the number of students in the school-grade-subject-year cell. Classrooms in which two-year leave-out value-added estimates cannot be constructed are excluded. Column 1 divides the data available for a teacher's value-added into information from when a teacher was in the same school and information from all other schools. Column 2 divides the data available for a teacher's value-added into information from when a teacher was in the same school, a similar school, and all other schools. Similar schools are defined as schools with mean test scores within 0.1 standard deviation units. The third-to-last row presents the p-value from an F-test of all coefficients in the corresponding column being jointly equal to 1.

Table 7. Differences in Teacher Quality Across Students and Schools

|	b / (se)	b / (se)	b / (se)	b / (se)	b / (se)
Lagged Test Score	0.024	0.013			
	(0.001)	(0.001)			
African-American			-0.030	-0.010	
			(0.004)	(0.001)	
Asian			-0.013	0.005	
			(0.003)	(0.001)	
Hispanic			-0.043	-0.010	
			(0.003)	(0.001)	
School Fraction African-American					-0.073
					(0.015)
School Fraction Asian					-0.076
					(0.025)
School Fraction Hispanic					-0.109
					(0.011)
School Fixed Effects		Yes		Yes	
Mean School FRPL Quartile					
R-sq	0.017	0.133	0.006	0.129	0.011
N	2,794,818	2,794,818	2,794,818	2,794,818	2,794,818

Notes: Each column reports coefficients and standard errors (in parentheses) from an OLS regression, where the dependent variable is the estimated teacher value-added using our baseline model (see Table 2, column 1). Standard errors are clustered at the teacher level. The regressions are estimated in a dataset at the student-subject-year level.

Table 8. Differences in Teacher Quality Across Students and Schools, Accounting for Teacher Experience

Panel A: Dependent Variable is Teacher Experience * Experience Coefficient

	b / (se)	b / (se)	b / (se)	b / (se)	b / (se)
Lagged Test Score	0.017	0.004			
	(0.003)	(0.002)			
African-American			-0.039	-0.005	
			(0.011)	(0.004)	
Asian			0.018	0.007	
			(0.009)	(0.003)	
Hispanic			-0.018	-0.000	
			(0.011)	(0.004)	
School Fraction African-American					-0.092
					(0.046)
School Fraction Asian					0.188
					(0.077)
School Fraction Hispanic					-0.015
					(0.035)
School Fixed Effects		Yes		Yes	
R-sq	0.001	0.343	0.001	0.343	0.002
N	2,897,425	2,897,425	2,897,425	2,897,425	2,897,425

Panel B: Dependent Variable is (Teacher VA with Experience Controls) + (Experience * Experience Coefficient)

|	b / (se)	b / (se)	b / (se)	b / (se)	b / (se)
Lagged Test Score	0.042	0.016			
	(0.003)	(0.002)			
African-American			-0.069	-0.015	
			(0.012)	(0.004)	
Asian			0.005	0.012	
			(0.010)	(0.003)	
Hispanic			-0.063	-0.010	
			(0.011)	(0.004)	
School Fraction African-American					-0.161
					(0.050)
School Fraction Asian					0.106
					(0.082)
School Fraction Hispanic					-0.131
					(0.038)
School Fixed Effects		Yes		Yes	
R-sq	0.006	0.338	0.002	0.337	0.005
N	2,689,580	2,689,580	2,689,580	2,689,580	2,689,580

Notes: Each column reports coefficients and standard errors (in parentheses) from an OLS regression. Standard errors are clustered at the teacher level. The regressions are estimated in a dataset at the student-subject-year level. In Panel A, the dependent variable is Teacher Experience * Experience Coefficient. We obtain the relevant experience coefficient by re-specifying equation (1) with controls for teaching experience. In Panel B, the dependent variable is the sum of teacher value-added (controlling for teacher experience) and the experience effects from Panel A.

Figure 1.
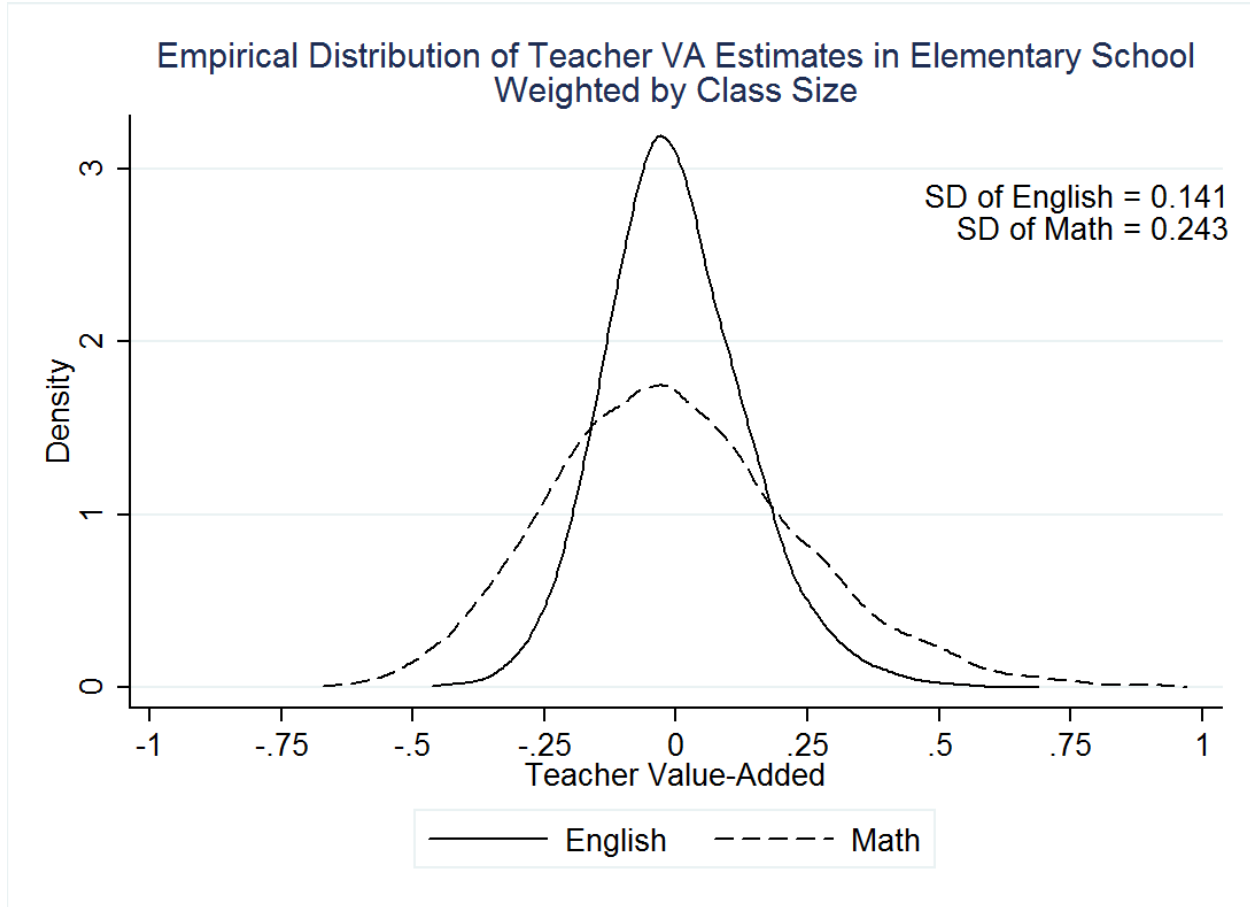
Empirical Distribution of Teacher VA Estimates in Elementary School Weighted by Class Size

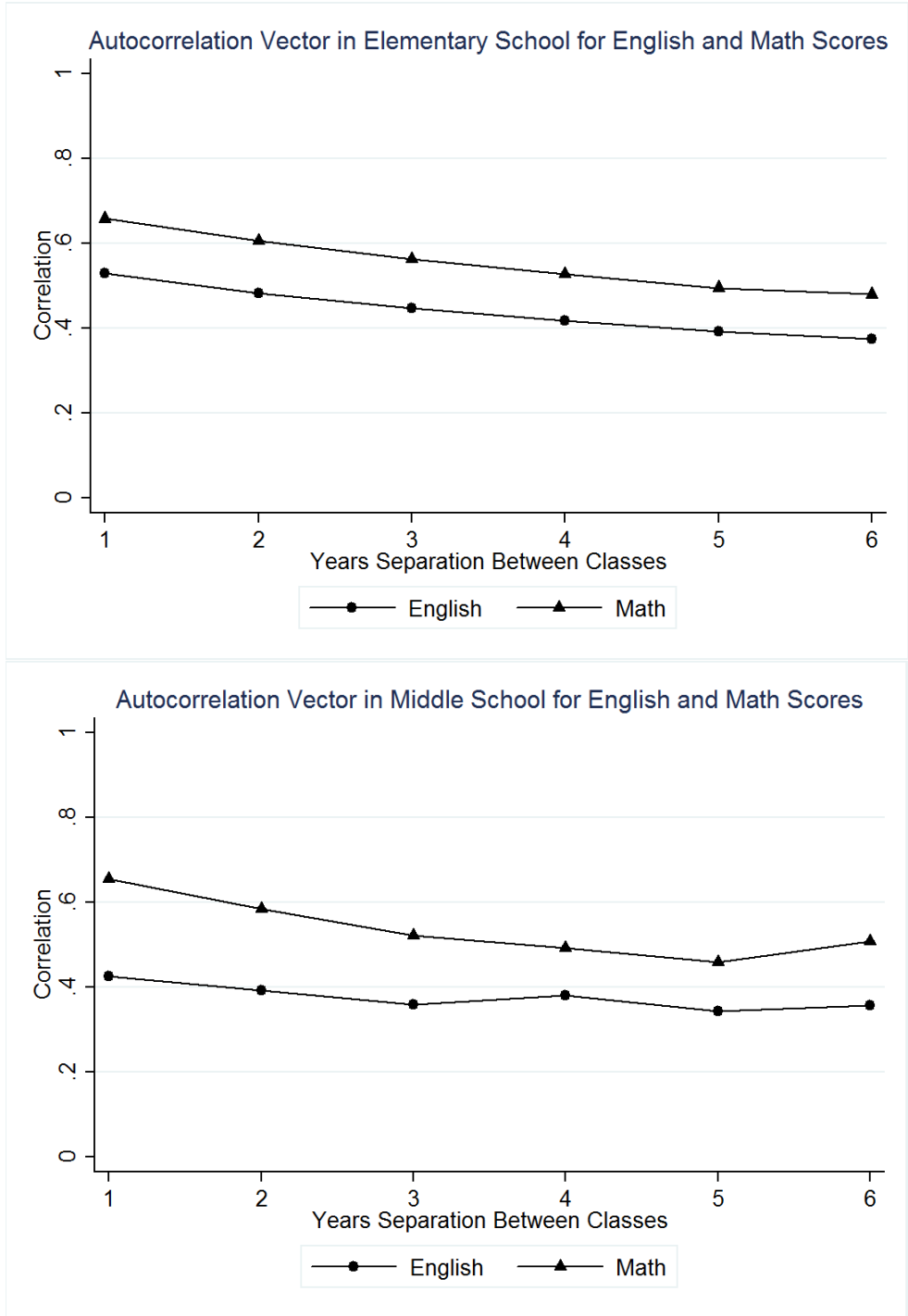SD of English = 0.141
SD of Math = 0.243

Notes: This figure plots kernel densities of the empirical distribution of predicted teacher effects for elementary school teachers. The densities are weighted by class size and are estimated using the Epanechnikov kernel with a bandwidth of .03. We also report the standard deviations of these empirical distributions of VA estimates, which are shrunken toward the mean to account for noise.

Figure 2.
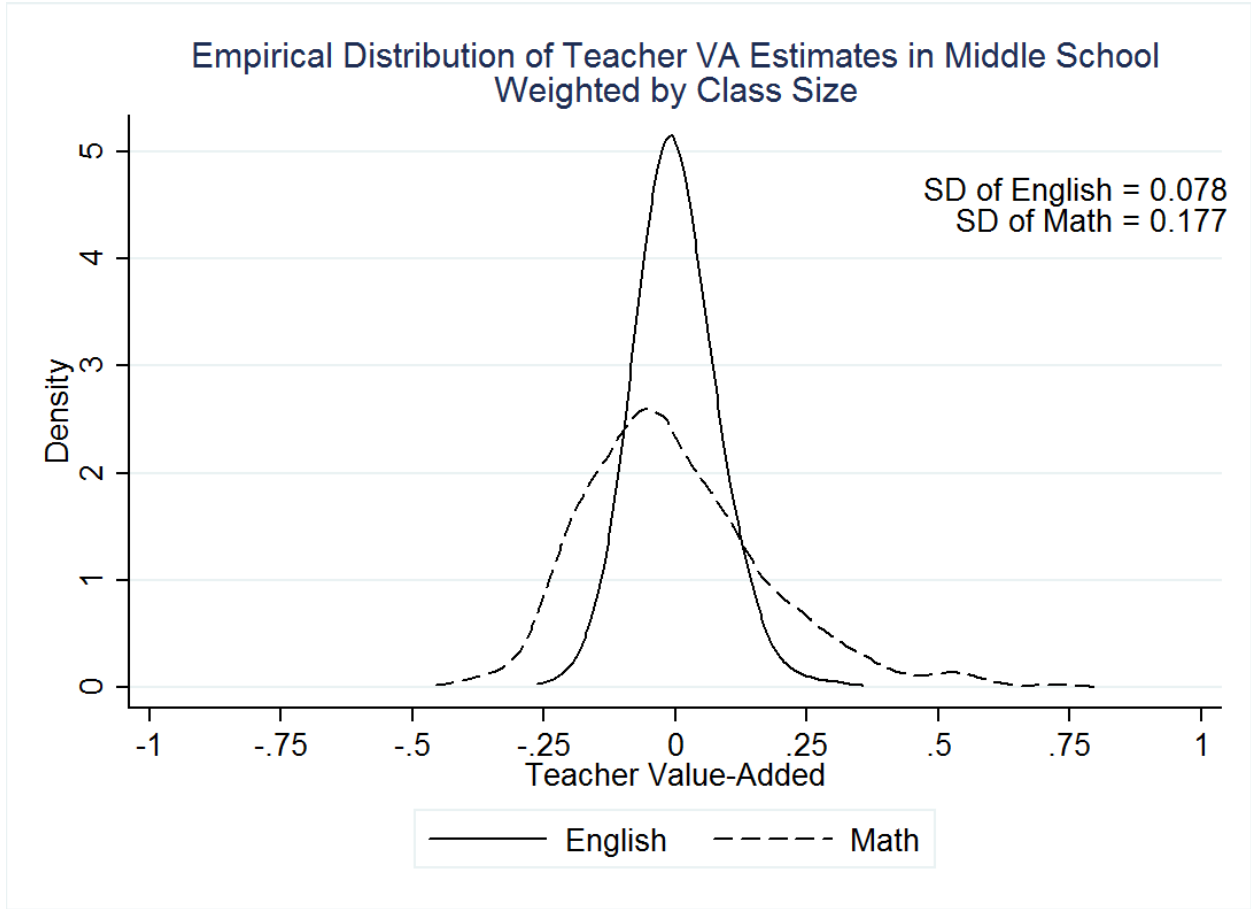
Empirical Distribution of Teacher VA Estimates in Middle School
Weighted by Class Size

SD of English = 0.078
SD of Math = 0.177

Notes: This figure plots kernel densities of the empirical distribution of predicted teacher effects for middle school teachers. The densities are weighted by class size and are estimated using the Epanechnikov kernel with a bandwidth of .03. We also report the standard deviations of these empirical distributions of VA estimates, which are shrunken toward the mean to account for noise.

Figure 3.

Autocorrelation Vector in Elementary School for English and Math Scores

Autocorrelation Vector in Middle School for English and Math Scores

Notes: These figures show the correlation between mean test-score residuals across classes taught by the same teacher indifferent years. The top panel plots autocorrelation vectors and the bottom panel plots autocorrelation vectors for middle school. See CFR 2014a Appendix A for details on this estimation procedure.

Figure 4.
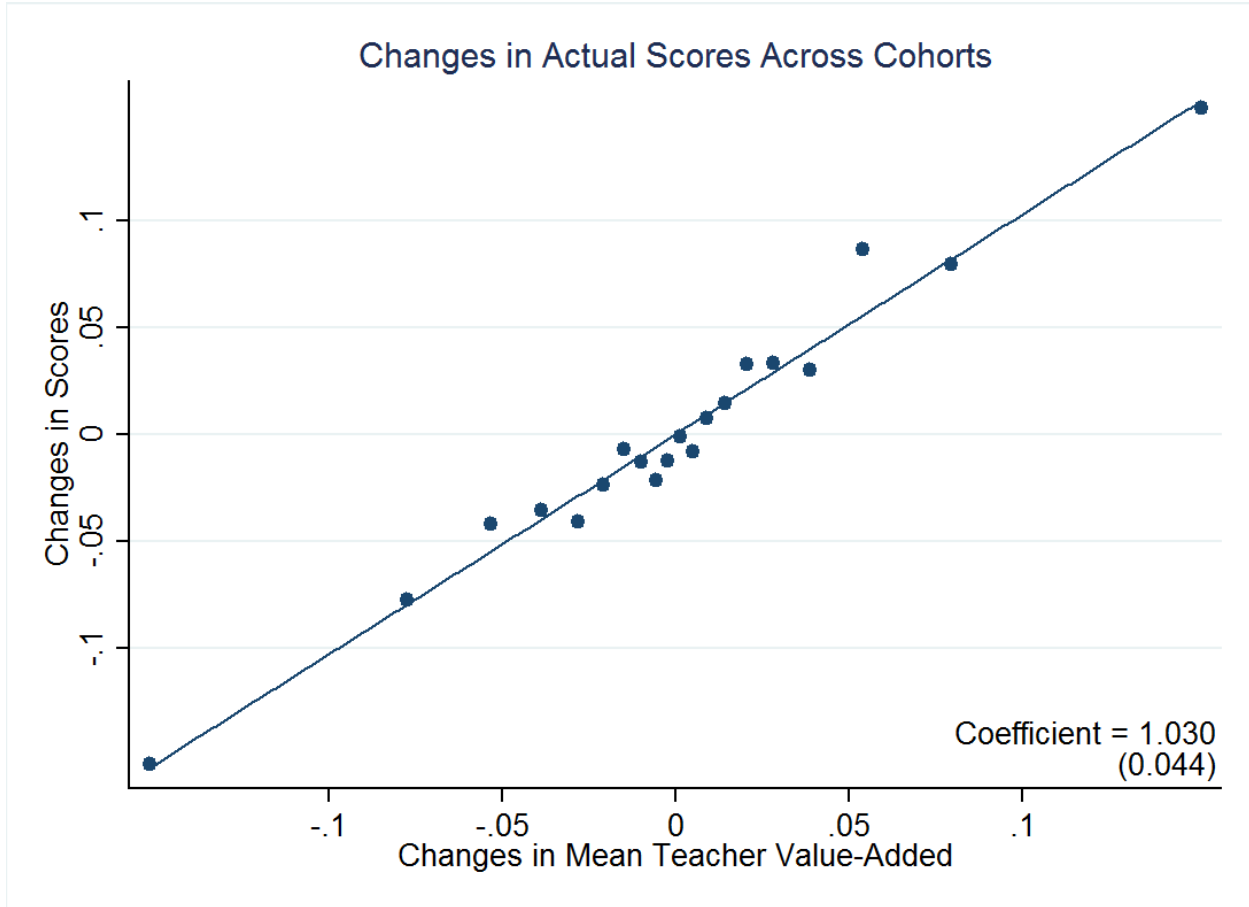
Notes: This figure presents a binned scatter plot and fitted line of changes in mean actual test scores versus changes in mean teacher predicted changes in VA across cohorts, which corresponds to the regression in column 1 of Table 2 (see Table 2 for details on the model). To construct these binned scatter plots, we follow the procedure detailed in CFR of first demeaning both the x- and y-axis variables by school year to eliminate any secular time trends. We then divide the observations into vingtiles based on their change in mean VA and plot the means of the y variable within each bin against the mean change in VA within each bin, weighting by the number of students in each school-grade-subject-year cell. The solid line shows the best linear fit estimated on the underlying micro data, which is presented in Table 2, column 1.