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.

[1] We thank Raj Chetty for helpful discussions and comments, and for providing the code used in the CFR study.
[2] Thomas J. Kane served as an expert witness for Gibson, Dunn, and Crutcher LLP to testify in *Vergara v. California*.

Introduction

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

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 sc

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 Δg in scores in a given grade and school from t-1 to t based on Δg 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

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

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,

$$(1)\quad \ddot{O}_{\text{خث}} \text{ أسي خث } \ddot{W}_{\text{خث}} \text{ اطى خث} + \text{ طى ج } | \text{ ج} + \text{خجث.}$$

$\ddot{O}_{\text{خث}}$ represents student i's test score in year t (standardized to have a mean of zero and standard deviation of one), $\ddot{W}_{\text{خث}}$ 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, $|_{\text{ج}}$ 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) ö أطي خثö سي خث

the weighted average of the residuals from years other than year t, with the weights derived from the drift parameters, [illegible]:

(4) [illegible]

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 8^{th} 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 Assignments: 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: 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

.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

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

$$(5)\ \Delta\bar{y}_{jgt} = \alpha + \beta\Delta\bar{Y}_{jgt} + \varepsilon_{jgt}$$

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, $\Delta\bar{Y}_{jgt}$. Then we calculate the average change in actual scores in each of the 20 groups, $\Delta\bar{y}_{jgt}$. 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 β 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 β 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 $\bar{\nu}^{-9}_{jt}$, we can use teaching experience to impute value-added for those with missing $\bar{\nu}^{-9}_{jt}$. 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 other subject 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 $\Delta\tilde{\upsilon}_{خثج}$ 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 *if* 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

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 $\Delta\ddot{o}_{\text{خ}\text{ث}\text{خ}}$ is the dependent variable. Rothstein reintroduces the problem when he uses $\Delta\ddot{o}_{\text{خ}\text{ث}\text{خ}٩٥}$ 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 $\Delta\tilde{u}_{\text{خ}\text{ث}\text{خ}}$ and $\Delta\ddot{o}_{\text{خ}\text{ث}\text{خ}٩٥}$. 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, $\Delta\tilde{u}_{\text{خ}\text{ث}\text{خ}}$, with a range of different specifications. For the specifications in the top row, the dependent variable is change in end-of-year achievement, $\Delta\ddot{o}_{\text{خ}\text{ث}\text{خ}}$; in the bottom row, the dependent variable is the change in lagged score (or baseline score), $\Delta\ddot{o}_{\text{خ}\text{ث}\text{خ}٩٥}$. Across all of our specifications, we find that the coefficient is stable and indistinguishable from one when change in end-of-year achievement, $\Delta\ddot{o}_{\text{خ}\text{ث}\text{خ}}$, is the dependent variable. However, we find that the coefficient is sensitive to the model specification when change in lagged achievement, $\Delta\ddot{o}_{\text{خ}\text{ث}\text{خ}٩٥}$, 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

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

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 eTS8cy greeal twhich a differengroupson or

studen ac

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,

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.

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

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

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:

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.

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

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

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

	Baseline Model (Two-Year Leave-Out)	Baseline Model with Three-Year Leave-Out		No Followers (IV)		No Within-School Movers (IV)
Dependent Variable:	**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) tews

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)

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.

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.

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.

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