Revisiting the Impacts of Teachers

Jesse Rothstein*

October 2014

Abstract

Using teacher switching as a quasi-experiment, Chetty, Friedman, and Rockoff (hereafter CFR) find that value added (VA) estimates of teacher effectiveness are not meaningfully biased by student sorting and are strongly correlated with students' later outcomes (CFR 2014a; 2014b). I successfully replicate CFR's key results in a new sample. Further investigation, however, reveals that the quasi-experiment is invalid: Teacher switching is correlated with changes in student preparedness. Estimates that adjust for this indicate moderate bias in VA scores. The association between VA and long-run outcomes is not robust and quite sensitive to controls.

*Goldman School of Public Policy and Department of Economics, University of California, Berkeley. E-mail: rothstein@berkeley.edu. I am grateful to Julien Lafortune for excellent research assistance and the North Carolina Education Research Data Center for access to data. I thank UC Berkeley labor seminar participants for helpful comments and David Card, Hilary Hoynes, Brian Jacob, Pat Kline, Diane Schanzenbach, Doug Staiger, Chris Walters, and especially Raj Chetty, John Friedman, and Jonah Rockoff for helpful discussions, though none are responsible for the results.

Value-added (hereafter, VA) models attempt to disentangle teacher effectiveness – defined as the teacher's causal effect on his or her students' test scores – from other factors that influence student achievement. They do so by controlling for students' prior-year scores and for other observed factors (e.g., free lunch status); any variation in student scores that is left after removing the components that are statistically attributable to observable differences in students is attributed to the teacher.

VA scores are used increasingly for high-stakes teacher evaluations, but this is highly controversial. One important criticism is that VA models may not successfully isolate teachers' causal effects on their students.[1] A student's classroom assignment may depend on parental requests or on teacher specializations that are not typically recorded in administrative databases, and it is not clear a priori whether the variables controlled in VA models are sufficient to absorb the endogeneity that this creates. If they are not, a teacher's VA score will reflect not just her effectiveness but also the types of students who were assigned to her. Teachers who have unusual assignments – e.g., those who are thought to be particularly effective with late readers, with hyperactive children, or with advanced students – may be rewarded or punished for this under VA-based evaluations, with potentially serious unintended consequences.

Rothstein (2010) finds that classroom assignments are significantly correlated with student characteristics – in particular, the student's full test score history beyond the prior year's score – that are predictive of later achievement and not typically controlled for in VA models. This implies that VA scores are biased. But it has not been possible to rule out the hypothesis that the magnitude of the bias is small enough to be ignorable (Rothstein, 2009; Guarino, Reckase, and Wooldridge, 2012).

Chetty, Friedman, and Rockoff (2014a; hereafter, CFR-I) introduce a new strategy for quantifying bias in VA models from non-random student assign-

[1] Critics have also argued that teacher VA is too noisily measured to be useful, that it misses teachers' effects on dimensions of achievement not captured by test scores, that VA-based evaluation creates incentives to target the measure (e.g., by teaching to the test) rather than to improve true achievement, and that the use of individual performance measures will undermine teacher cooperation.

ments. They examine teacher switches – events where one teacher leaves a school and is replaced by another. If VA scores are unbiased, the year-over-year change in school-level scores, averaging over both students in the switching teachers' classes and others taught by teachers who remained in the school, should be consistent with a prediction based on the difference between the departing and arriving teachers' VA scores.[2] But if VA scores are biased by non-random student assignments, the change in student outcomes will generally be smaller than the VA-based prediction. The test is formalized in CFR-I and in Section 1 below.

Using data from a large, unnamed school district, CFR-I find that VA scores successfully predict changes in test scores following teacher switching, and conclude that biases in teachers' VA scores are minimal or nonexistent. A companion paper, CFR (2014b; hereafter CFR-II), finds that teacher VA is associated with students' long-term outcomes, such as their earnings as young adults, both in cross-sectional comparisons and in quasi-experimental analyses of teacher switching. CFR thus conclude that high-VA teachers have large effects on students' long-run outcomes.

I replicate CFR's analyses using a statewide dataset from North Carolina. Section 2 describes the North Carolina data, and Section 3 presents replication results. I successfully reproduce all of CFR-I's key results.

In the North Carolina data, as in CFR's district, one of CFR's robustness tests yields results that appear inconsistent with those of the main specification. To understand this, I investigate the validity of the teacher switching quasi-experiment, in Section 4. Like all quasi-experiments, this one relies on an assumption that the treatment – here, teacher switching – is as good as random. I find that it is not: Teacher switching is correlated with changes in students' prior-year scores. Exiting teachers tend to be replaced by teachers with higher measured VA when students' prior achievement is increasing for other reasons, and by teachers with lower measured VA when student pre-

[2]The CFR-I analysis, and my replication, is actually conducted at the school-grade-subject level, and some teacher switches occur when a teacher merely changes grades within the same school.

paredness is declining. CFR have confirmed that this result holds in their sample as well (CFR 2014c).[3]

The evidence that the teacher switching "treatment" is not randomly assigned implies that CFR-I's quasi-experimental analyses, which do not control for changes in student preparedness, cannot be interpreted causally. I thus turn, in Section 5, to specifications that augment CFR-I's by adding controls for changes in students' prior-year scores. If teacher switching is random conditional on this observable, these specifications identify the prediction bias coefficient of interest. Results indicate that VA scores over-predict the change in student learning, with a prediction coefficient between 0.8 and 0.9. This implies that the bias component of VA scores is statistically and practically significant, with a magnitude squarely in the middle of the range identified as plausible by Rothstein's (2009) simulations.

Section 6 turns to CFR-II's analyses of teachers' long-run effects. Again, I successfully replicate all of the key results, albeit using a different set of long-run outcome measures that are available in the North Carolina data.

It is not clear that the association between VA and long-run outcomes can be interpreted causally. The evidence of bias in VA scores means that the association between a teacher's VA and students' long-run outcomes may reflect the student sorting component of the VA score rather than the teacher's true effect. Moreover, even if this issue is set aside there is still a concern that students assigned to high-VA teachers may be advantaged in ways that are predictive of the students' long-run outcomes, implying that the estimated "effect" of being assigned to a teacher with high estimated VA is upward biased. In both CFR's district and the North Carolina sample, teachers' measured VA is correlated with students' prior scores and other observables.

Neither CFR-II's observational estimates nor their quasi-experimental estimates of teachers' long-run effects control fully for students' observed, predetermined characteristics. Unfortunately, the North Carolina data do not

[3]CFR (2014c) attribute the result to "mechanical" effects deriving from the use of the overlapping data for the VA calculations and for measuring prior achievement. I investigate this explanation in Appendix B, and find that it cannot account for the results.

provide detailed measures of family economic status, so I am limited in my ability to explore the sensitivity of the results. Nevertheless, when I control for the observables that are available, I estimate effects of high-VA teachers on students' long-run that are much smaller than is implied by CFR-II's methods. CFR-II present evidence suggesting that additional controls, were they available, would further diminish the estimated effects. Even with my limited controls, however, quasi-experimental estimates are generally not significantly different from zero, and all point estimates are smaller than when the controls are omitted (as in CFR-II's analyses).

Both the bias estimates and the estimated long-run effects can be interpreted causally only under strong assumptions of selection-on-observables. At a minimum, one can conclude that analyses like those proposed by CFR-I and CFR-II do not provide strong evidence about either the magnitude of sorting biases or the effects of high-VA teachers on students' later outcomes, and that new research designs will be needed to provide credible estimates of either. What evidence there is suggests that VA scores are importantly biased by student sorting and that the long-run effects of having a teacher with a high (measured) VA score are substantially smaller than are implied by CFR-II's results.

1 The teacher switching quasi-experiment

I describe the quasi-experimental design briefly here, drawing on CFR-I and adopting their notation. Readers are referred to their paper for a more complete description.

1.1 Teacher value-added

The data generating process for student i's test score in year t, A^*_{it}, is

$$A^*_{it} = X_{it}\beta + \mu_{j(i,t)t} + \epsilon_{it}, \tag{1}$$

where X_{it} is a vector of observables, including the student's prior year score; $j(i,t)$ represents student i's teacher in year t; μ_{jt} is the causal effect of teacher j on her students; and ϵ_{it} is an unobserved shock that may be correlated among students in the same classroom.[4] CFR estimate β via a regression of A^* on X, controlling for teacher fixed effects, and compute the average residual score among students in class (j,t):

$$\bar{A}_{jt} = \frac{1}{n_{jt}} \sum_{i:\, j(i,t)=j} A^*_{it} - X_{it}\hat{\beta}. \tag{2}$$

β is very precisely estimated. If the within-teacher regression is unbiased, then to first approximation

$$\bar{A}_{jt} = \mu_{jt} + \bar{\epsilon}_{jt}, \tag{3}$$

where $\bar{\epsilon}_{jt} \equiv \frac{1}{n_{jt}} \sum_{i:\, j(i,t)=j} \epsilon_{it}$.[5]

CFR's primary VA measure is the linear forecast of the teacher's causal effect in year t, μ_{jt}, given her students' average residuals in other years $s \neq t$. The forecast is:

$$\hat{\mu}_{jt} = \sum_{\tau \in T_j, \tau \neq 0} \psi^{T_j}_{|\tau|} \bar{A}_{j,t+\tau}, \tag{4}$$

where $T_j \equiv \left\{\tau | \bar{A}_{j,t+\tau} \text{ is observed for teacher } j\right\}$ and ψ^{T_j} is a vector of best-linear-predictor coefficients.[6] These coefficients vary with $|\tau|$ and with the set T_j – when a teacher is observed for many years, the prediction will put little weight on any one of them, but when T_j is smaller the available data will be weighted more heavily.[7] Both in theory – given CFR-I's stationarity assump-

[4]CFR use data from tests in math and reading, treating a student's scores on the two tests as two different observations and interacting most coefficients with a subject indicator. I suppress subject subscripts here for readability.

[5]CFR-I do not state assumptions under which $\hat{\beta}$ is unbiased. One can imagine decomposing μ_{jt} into a permanent teacher-level component and a transitory component, $\mu_{jt} = \tilde{\mu}_j + u_{jt}$. CFR-I's procedure allows $\tilde{\mu}_j$ to be correlated with $\bar{\bar{X}}_j \equiv E\left[X_{it} \,|\, j(i,t) = j\right]$ across teachers but implicitly assumes that u_{jt} is uncorrelated with $\bar{X}_{jt} \equiv E\left[X_{it} \,|\, j\,(i,t) = j,\, t\right]$ across classes within teachers.

[6]CFR refer to $\hat{\mu}_{jt}$ as teacher j's value-added. That term is more commonly used for the teacher's true causal effect μ_{jt}. In this paper, I reserve "value added" for μ, and refer to $\hat{\mu}$ as the "predicted" or "forecast" value added.

[7]CFR-I assume that the within-teacher, between-year covariance of $\bar{A}_{jt}$ depends only

tions – and empirically, the sum of coefficients for any teacher, $\sum_{\tau \in T_j, \tau \neq 0} \psi^{T_j}_{|\tau|}$, is less than one, so $\hat{\mu}_{jt}$ is "shrunken" relative to a weighted average of the $\bar{A}_{j,t+\tau}$s.

1.2 The teacher-switching quasi-experiment

A central question in the VA literature (see, e.g., Rothstein 2009, 2010) is whether $\bar{A}_{jt}$ provides an unbiased estimate of teacher j's causal effect, or whether it is biased by sorting of students to teachers on the basis of characteristics that are not controlled in the VA model. CFR-I distinguish two kinds of bias, which they call "forecast bias" and "teacher-level bias," and argue that the former is more relevant to policy. VA scores are forecast-unbiased if $E\left[\mu_{jt}|\hat{\mu}_{jt}\right] = \hat{\mu}_{jt}$.

Importantly, forecast bias cannot be identified by comparing $\bar{A}_{jt}$ to $\hat{\mu}_{jt}$ in observational data, as the coefficients ψ are chosen precisely to ensure that the linear projection of $\bar{A}_{jt}$ onto $\hat{\mu}_{jt}$ (which, recall, is estimated only using data on the teacher's students' scores in years other than t) has a slope of one, even if both $\bar{A}_{jt}$ and $\hat{\mu}_{jt}$ are biased relative to μ_{jt}. Testing for forecast bias requires a strategy for constructing a proxy for μ_{jt} that is unbiased, or at least subject to a *different* form of bias than is $\bar{A}_{jt}$.

Kane and Staiger (2008) develop an experimental test. Starting with pairs of teachers A and B teaching in the same school and grade, they randomly assign students among the two teachers in the pair. This ensures that any student sorting component of the between-teacher difference in average scores in the year of random assignment, $\bar{A}^*_{At} - \bar{A}^*_{Bt}$, is uncorrelated with the sorting component of the between-pair difference in average predicted value-added based on prior years' data, $\hat{\mu}_{At} - \hat{\mu}_{Bt}$. Thus, the regression of the former on the latter identifies the coefficient of a regression of $\mu_A - \mu_B$ on $\hat{\mu}_{At} - \hat{\mu}_{Bt}$, so should have a coefficient of one if the VA measures are forecast-biased and a

on the number of elapsed years: $Cov\left(\bar{A}_{jt}, \bar{A}_{js}\right) = \sigma_{|t-s|}$. They use data for all pairs of observations $|t-s|$ years apart to estimate $\sigma_{|t-s|}$, then use the vector of σs to construct the best predictor coefficients ψ^T for each set T.

coefficient less than one if they are forecast-biased.[8] Unfortunately, Kane and Staiger's sample was small and non-representative (see also Kane, McCaffrey, Miller, and Staiger 2013; Rothstein and Mathis 2013), so results were not decisive.

CFR-I extend Kane and Staiger's experimental design to examine "quasi-experiments" created when a teacher leaves a school (or switches grades within a school) and is replaced by another. If VA scores are forecast-unbiased, the change in average predicted VA among the teachers in the school-grade cell should accurately predict the change in average student test scores. CFR define Q_{sgt} as the average of $\hat{\mu}_{jt}$ among all teachers j in grade g at school s in year t, weighted by the number of students taught.[9] The predicted change in teacher impacts is $\Delta Q_{sgt} = Q_{sgt} - Q_{sg,t-1}$. If we let $\bar{\mu}_{jgt}$ represent the analogous weighted average of teachers' true causal effects μ_{jt} and $\Delta\bar{\mu}_{jgt}$ its difference, forecast-unbiasedness implies that $E\left[\Delta\bar{\mu}_{st}|\Delta Q_{sgt}\right] = \Delta Q_{sgt}$.

Of course, $\Delta\bar{\mu}_{jgt}$ is not directly observed. CFR-I use the change in average student scores, $\Delta\bar{A}^{*}_{sgt}$, in place of it. Importantly, the process by which students are sorted across schools is distinct from the process of within-school sorting to teachers, which might bias ΔQ_{sgt} but should have no effect on $\Delta\bar{A}^{*}_{sgt}$. Thus, so long as changes in teacher quality are uncorrelated with changes in the school population, the coefficient of a regression of $\Delta\bar{A}^{*}_{sgt}$ on ΔQ_{sgt} should have a coefficient of one if VA scores are unbiased by within-grade student sorting, and a coefficient less than one if they are biased.

1.3 Evaluating the quasi-experiment

All quasi-experiments rely on untestable assumptions that the "treatment" is uncorrelated with other determinants of the outcome. Here, the treatment is

[8] This test requires the true VA of teacher A to be uncorrelated with that of teacher B, and the measurement error in the two teachers' VA estimates to be similarly uncorrelated. Without this, the difference of unbiased predictors need not be an unbiased predictor of the difference (and vice versa). CFR-I's test relies on a similar assumption that the only available information that is informative about a teacher's causal effect is the sequence of average residuals for that teacher's students in other years.

[9] As noted above, CFR stack math and reading observations, so the actual level of observation is school-grade-subject-year.

the change in estimated teacher quality, ΔQ_{sgt}. CFR-I argue that there is no reason to expect that the arrival or departure of a teacher with high (or low) $\hat{\mu}_{jt}$ would be correlated with changes in other determinants of average student outcomes at the school-grade level; that is, that

$$\begin{aligned} cov\left(\Delta Q_{sgt}, \Delta \bar{A}^*_{sgt}\right) &= cov\left(\Delta Q_{sgt}, \Delta \bar{\mu}_{sgt} + \Delta \bar{X}_{sgt}\beta + \Delta \bar{\epsilon}_{sgt}\right) \\ &= cov\left(\Delta Q_{sgt}, \Delta \bar{\mu}_{sgt}\right). \end{aligned} \tag{5}$$

The assumption that $cov\left(\Delta Q_{sgt}, \Delta \bar{\epsilon}_{sgt}\right) = 0$ is untestable, as $\Delta \bar{\epsilon}_{sgt}$ is unobserved. But it is possible to test whether $cov\left(\Delta Q_{sgt}, \Delta \bar{X}_{sgt}\beta\right) = 0$.[10] Violations of this restriction would indicate that teacher-switching is not a valid quasi-experiment and that a simple regression of $\Delta \bar{A}^*_{sgt}$ on ΔQ_{sgt} does not estimate the projection coefficient of interest. At a minimum, one would want to control for $\Delta \bar{X}_{sgt}$ in this regression; one might also question the validity of the untestable assumption $cov\left(\Delta Q_{sgt}, \Delta \bar{\epsilon}_{sgt}\right) = 0$ and hesitate to attach a causal interpretation to even the regression with controls.

The most important component of X_{it} is the student's score in year $t-1$. Below, I present regressions of the change in mean lagged scores at the school-grade-year level on ΔQ_{sgt} as a test of the identifying assumption of the quasi-experiment.

1.4 Missing data and Empirical Bayes predictions

CFR-I construct $\hat{\mu}_{jt}$ and $\hat{\mu}_{j,t-1}$ only from $\bar{A}_{jt'}$ observations from years other than $t-1$ and t. A problem arises when a teacher is observed only in those years. For these teachers, CFR-I set $\hat{\mu}_{jt}$ and $\hat{\mu}_{j,t-1}$ to missing, excluding these teachers from their right-hand-side variable ΔQ_{sgt}. They also exclude these teacher's students from the left-hand-side variable $\Delta \bar{A}^*_{sgt}$. They present an al-

[10] $\hat{\mu}_{jt}$ is constructed from residuals of regressions of test scores on X_{it}. But ΔQ_{sgt} need not be orthogonal to $\Delta \bar{X}_{sgt}\beta$. Recall that CFR-I estimate β using within-teacher variation, so between-teacher variation could produce a non-zero covariance between $\bar{X}_{jt}$ and $\bar{A}_{jt}$. In addition, the use of leave-one-out EB predictions can generate an association between $\hat{\mu}_{jt}$ and $\bar{X}_{jt}$ even if $cov\left(\bar{A}_{jt}, \bar{X}_{jt}\right) = 0$. Finally, the differenced, aggregated variables can covary even if the teacher-by-year versions do not.

ternative specification that includes the teachers and their students, assigning predicted VA of zero to the teachers. This yields importantly different results, both in their tables and in my replication sample. It is thus worth considering the logic of the two approaches.

As CFR-I note, their predicted VA $\hat{\mu}_{jt}$ can be seen alternately as a "shrinkage" estimator that pulls noisy signals of a teacher's impact in other years $s \neq t$ toward zero in inverse proportion to the signal-to-noise ratio of the aggregated signals; as a best linear predictor of μ_{jt}; as the posterior mean of μ_{jt} given data $\left\{\bar{A}_{js}\right\}_{s \neq t}$; or as an Empirical Bayes estimate of μ_{jt}.

An important property of $\hat{\mu}_{jt}$ can be best understood in the special case where μ_{jt} is assumed to be constant across t within teacher and $\bar{\epsilon}_{jt}$ is i.i.d. across t, as in an earlier version of CFR's study (CFR 2011) and in the earlier work on which they draw (Kane and Staiger, 2008). Then the weights $\psi_{\tau}^{T_j}$ are uniform across τ – as much weight is put on an observation from ten years ago as is put on last year's measure.[11] In this case, $\hat{\mu}_{jt}$ can be written as a simple average of the teacher's students' mean residual scores in all other years $s \neq t$, multiplied by a shrinkage factor ψ_j. This factor, defined in CFR-I's equation (9), is the reliability of the average of residual scores, seen as a noisy measure of μ_j. Thus, $0 < \psi_j < 1$. It approaches one as the number of years of available data rises toward infinity, and approaches zero as the amount of data shrinks toward zero. Intuitively, ψ_j shrinks the teacher's observed performance toward the grand mean of μ, 0, allowing teachers to deviate substantially from the grand mean only when their posted performance is quite different from the average and when enough data is available to ensure that this is not a fluke.

Now consider the problem of predicting the VA of a teacher who appears in the data only in the $\{t-1, t\}$ window. The mean residual score – which only uses data from years outside that window – cannot be constructed. If VA estimation is seen as a signal extraction problem, the signal for this teacher has zero precision, implying complete shrinkage, or $\psi_j = 0$. Under shrinkage, best prediction, or Empirical Bayes interpretations, this teacher's VA should

[11] Following CFR-I, I abstract from differences in class size across years. See Kane and Staiger (2008) for a clear exposition that accounts for this.

be set equal to the grand mean: $\hat{\mu}_{jt} = E[\mu] = 0$. Assigning this value is nothing more than treating the teacher the same way as other teachers are treated.[12]

CFR (2011) found evidence that μ_{jt} is *not* stable across t, and subsequent versions of the paper implement a new EB estimator designed to allow for "drift" in μ_{jt} (see equation 4 above). But the shrinkage interpretation is still valid. CFR-I's $\hat{\mu}_{jt}$ can be seen as a *weighted* average of $\bar{A}_{js}$, $s \neq t$, with weights declining in $|s-t|$, again multiplied by a shrinkage factor ψ_{jt} that corresponds to the reliability of the weighted average as a signal of μ_{jt}.[13] As in the no drift case, as the amount of data from other years shrinks to zero, so does ψ_{jt} and $\hat{\mu}_{jt}$. A teacher with no data from other years should be assigned $\hat{\mu}_{jt} = 0$.

What of CFR-I's preferred strategy of excluding classrooms of teachers observed only once or twice? The exclusion of these teachers' predicted VA scores from Q_{sgt} is defensible if $\hat{\mu}_{jt}$ is missing at random. This is a strong assumption. But CFR-I's procedure requires an even stronger restriction, as they exclude these teachers' students from the calculation of $\bar{A}^*_{sgt}$ as well. This requires that mean scores in the excluded classrooms equal mean scores in the included classrooms conditional on the average predicted VA of the included teachers. I show below that this is unlikely given the overall structure of the data, and that the $\Delta\bar{A}^*_{sgt}$ computed only from classrooms included in CFR-I's Q_{sgt} calculation is biased relative to the whole-school $\Delta\bar{A}^*_{sgt}$ in a way that is correlated with CFR-I's ΔQ_{sgt}, skewing the quasi-experimental test statistic.

[12] This assumes that nothing is known about a teacher other than her sequence of mean student residuals $\bar{A}_{jt}$. One could adopt a richer model, relying on a vector of teacher observables Z_{jt} (e.g., teacher experience) that help to predict μ_{jt}. Then EB VA predictions for teachers observed many times shrink the teachers' average student residuals toward $E[\mu_{jt}|Z_{jt}]$, while the EB prediction for a teacher observed only in the $\{t-1, t\}$ window is $\hat{\mu}_{jt} = E[\mu_{jt}|Z_{jt}]$. CFR-I eschew the use of other information to predict μ_{jt} (see CFR 2014a, footnote 19).

[13] Although CFR-I do not present their prediction coefficients ψ_τ this way, their equation (6) and the surrounding discussion imply that the sum of ψ_τ over all $\tau \in T_j$ is less than one, and smaller the fewer years are available.

2 North Carolina data

I draw on administrative data for all students in the North Carolina public schools in 1997-2011, obtained under a restricted-use license from the North Carolina Education Research Data Center. End-of-course scores in math and reading are available for students in grades 3 through 8. Third grade students are given "pre-tests" in the Fall; I treat these as grade 2 scores.[14] I standardize all scores within each year-grade-subject cell.

The North Carolina administrative records record the identity of the test proctor. This is usually but not always the student's regular classroom teacher, though in grades where students are taught by separate teachers for different subjects the proctor for the math test might be the English teacher. I thus limit the sample to students in grades 3-5, for whom classrooms are generally self-contained. I use data on teachers' course assignments to identify exam proctors who do not appear to be the regular classroom teacher.

Many studies using the North Carolina data exclude such proctors and their students. That is not feasible here, as the quasi-experimental strategy requires data on all students in the school-grade cell. As an alternative, I assign each such proctor a unique teacher code that does not match across years.[15] This ensures that student achievement data is not used to infer the proctoring teacher's impact.

Not all of CFR-I's covariates are available in the North Carolina data. In particular, I do not have measures of absences and suspensions, enrollment in honors classes, or foreign birth. Thus, the control variables in my X vector are a subset of those that CFR-I use: Cubic polynomials in prior scores in the same and the other subject, interacted with grade; gender; age; indicators for special education, limited English, grade repetition, year, grade, free lunch

[14] Pre-test scores are not available in 2007-2009.

[15] I use a somewhat less restrictive threshold for a valid assignment than in past work (e.g., Clotfelter, Ladd, and Vigdor, 2006; Rothstein, 2010), to maximize the number of teachers for whom VA scores can be computed. Insofar as I fail to identify teachers who merely proctored the exam, this will attenuate the within-teacher autocorrelation of $\bar{A}_{jt}$. My estimates of these autocorrelations are larger than those reported by CFR-I for their data. See Appendix Table 1.

status, race/ethnicity, and missing values of any of these; class- and school-year- means of the individual-level controls; cubics in class- and school-grade mean prior scores; and class size.

For my analysis of long-run impacts, I focus on five outcomes that can be measured in students' high school records: Whether the student graduated from high school; whether she stated on a high school exit survey that she planned to attend college after graduation; whether she planned specifically to attend a four-year college; her high school grade point average; and her high school class rank. These outcomes are more proximate than CFR-II's outcomes, which mostly measure post-high-school experiences. They also vary in their availability; I focus only on cohorts for which they are available for most students. Students who do not appear in the North Carolina high school records are excluded from this analysis, while those who drop out of high school are assigned as non-college-bound.

My sample consists of 8.6 million student-year-subject observations, spread across three grades, two subjects (math and reading), 1,724 schools, and 15 years. 6.3 million of these observations can be linked to 36,888 valid teachers; 5.6 million also have non-missing end-of-grade and prior-year test scores. This is a bit smaller than CFR-I's sample, which contains approximately 18 million student-year-subject observations, but I have valid teacher IDs for a larger share (73% vs. 58%). I have non-missing leave-one-out predicted VA scores for 240,660 teacher-year-subject cells, with an average of 22.4 students per cell.

3 Replication

I use CFR's Stata programs (Chetty, Friedman, and Rockoff, 2014d) to reproduce their VA calculations and analyses in the North Carolina data. Appendix A presents several of CFR's results and my replications in parallel. Summarizing briefly, math VA is more variable in North Carolina than in CFR's sample, while English VA is more similar. In both math and English, the autocorrelation of teacher VA across years is higher in the North Carolina data (Appendix Table 1), implying less noise in the measurement process and perhaps also less

drift in teachers' true VA.

In both samples, students with higher prior-year scores tend to be assigned to teachers with higher predicted VA (Appendix Table 2). I find that special education students get higher VA teachers, on average, where CFR found the opposite, but the magnitude is small. A bigger difference relates to school composition: CFR-I find that the school minority share is insignificantly correlated with average teacher VA, but I find a larger, significant, positive correlation in North Carolina.

Table 1 presents my replication of CFR-I's main quasi-experimental analysis (Panel A) along with corresponding estimates from CFR-I (Panel B). Column 1 presents coefficients from a regression of the year-over-year change in average scores at the school-grade-year level on the change in average predicted VA, with year fixed effects, as in CFR-I's Table 4, Column 1. I follow CFR-I's specification (using their code): The regression is estimated on school-grade-year-subject-level aggregates and weighted by the number of students in the school-grade-year cell; standard errors are clustered at the school-cohort level; and classrooms for which the teacher's predicted VA cannot be constructed (because she is not seen in other years) are omitted from both the dependent and independent variables.[16] Column 2 repeats this specification with school-year fixed effects (as in CFR-I, Table 4, Column 2). In each case, the North Carolina results closely match those obtained by CFR-I: My point estimates are slightly larger than CFR-I's, and none indicate an effect significantly less than one.

Panel A of Figure 1 presents a binned scatter plot that illustrates the Column 2 specification. School-grade-subject-year observations are divided into twenty bins by the change in average predicted VA (i.e., by ΔQ_{sgt}); the Figure shows the average change in end-of-year students scores in each bin plotted against the average predicted change based on teachers' VA measures, after residualizing each against school-year fixed effects. The points are all

[16]Following CFR's code, the mean scores used for the dependent variable include classrooms taught by teachers observed in both $t-1$ and t but not in other years, even though these teachers are excluded from the independent variable.

quite close to the 45 degree line, and the slope is not significantly different from one.

Columns 3 and 4 of Table 1 present two of CFR's robustness checks. In Column 3, based on CFR-I, Table 4, Column 4, the dependent variable is the average *predicted* score, constructed as the fitted value from a regression of students' scores on parent characteristics. [17] In both samples, the year-on-year change in mean predicted VA is uncorrelated with the change in mean predicted scores. I show below, however, that changes in other predetermined student characteristics, not examined by CFR-I, are correlated with the change in teacher VA in the North Carolina sample.

Column 4 returns to the specification from Column 1, but adds classrooms with missing teacher VA predictions (16% of classrooms in the CFR-I sample and 32% in my North Carolina sample) to both the left and right hand side averages, with the teachers' predicted VA set to zero. This yields coefficients around 0.87, with confidence intervals that exclude 1 in both samples.

These estimates, taken on their face, indicate that the hypothesis of zero prediction bias is rejected. CFR-I instead attribute the result to measurement error in average teacher quality, induced by the imputation of zero predicted VA to previously excluded teachers, that leads to attenuation bias. It is well known that classical measurement error in independent variables attenuates regression coefficients, as it inflates the variance of the independent variable (which enters the demoninator of the regression coefficient) but not its covariance with the dependent variable (in the numerator). But the imputation of the grand mean to some teachers does not produce classical measurement error. In general, it reduces the variance of Q_{sgt} and (in likely data configurations) ΔQ_{sgt}. It is not clear that this kind of imputation leads to attenuation.

One way to assess the divergent results in Columns 1 and 4 of Table 1 is to consider separately the change in the independent and the dependent variables.[18] CFR-I's measurement error explanation implies that the change

[17] CFR-I's prediction is based on mother's age, marital status, parental income, 401(k) contributions, and homeownership, all drawn from tax files. Mine is based only on parental education, as reported in the North Carolina end-of-grade test score files through 2007.

[18] Another approach, pursued by CFR-I, is to restrict the analysis to the subsample of

in coefficients derives from the change in the independent variable. The final columns of Table 1 show that this is not correct. Column 5 uses all teachers to construct the independent variable, imputing zero predicted VA for those not observed in other years, but limits the sample used to construct the dependent variable to teachers whose VA is not imputed. Under CFR-I's explanation, the estimate in this column should be attenuated just as was the one in Column 4. But this is not the case – the coefficient is 1.205, significantly greater than one. Column 6 does the reverse, using all students to construct the dependent variable but excluding teachers with missing predicted VA from the independent variable. Here, the coefficient is only 0.659.[19]

Evidently the decline in the coefficient from Column 1 to Column 4 derives more from the change in the dependent variable than from the change in the independent variable. This is inconsistent with CFR-I's explanation. In the next Section, I show that the use of non-random subsets of students to construct the dependent variable introduces sample selection that is positively correlated with the measured change in average predicted VA, biasing the coefficients in Columns 1 and 2 upward relative to the parameter of interest.

---

school-grade-year cells for which leave-two-out predictions can be formed for all of the teachers. This is less than one-third of the total. In both data sets (see Appendix Table 5), coefficients estimated from this subsample are smaller but are statistically indistinguishable from one. Variation in the independent variable in the subsample comes disproportionately from teachers who have not switched schools. (ΔQ_{sgt} can vary without between-school mobility because teachers switch grades within schools or because the $t-1$ and t VA predictions put different weights on the teacher's performance in each other year.) In the North Carolina data, school switchers account for 71% of the variance of ΔQ_{sgt} in the full sample but only 57% in the complete-data sample. In any event, even in the restricted sample controlling for the change in lagged scores in a model for the change in end-of-grade scores, as discussed in Section 5, reduces the ΔQ_{sgt} coefficient substantially and the null hypothesis of zero prediction bias is rejected. See Appendix Table 5, Panel C.

[19]The qualitative results in Columns 4-6, Panel A, are robust to including school-year fixed effects and (for Column 4) to limiting the sample to school-grade-year cells included in the Column 1 sample.

4 Assessing the Validity of the Quasi-Experiment

A standard approach to testing the validity of an experiment or quasi-experiment is to estimate the correlation between supposedly randomly assigned treatment and pre-treatment covariates, particularly those that might be correlated with the outcome variable. Rothstein (2010) uses this method to assess teacher-level VA estimates, finding that students' teacher assignments are correlated with the students' test scores in earlier grades. The same approach can be applied to the teacher switching quasi-experiment: This research design relies on an assumption that the change in average predicted VA at the school-grade-year level, ΔQ_{sgt}, is as good as randomly assigned, so it should be uncorrelated with changes in student characteristics that are predictive of outcomes.

As discussed above (see Column 3 of Table 1), CFR-I report an exercise of this form. They find that the correlation between ΔQ_{sgt} and an index of parents' characteristics, weighted to best predict end-of-year scores, is essentially zero. But parents' permanent characteristics are unlikely to capture the dynamic sorting that Rothstein (2010) found to be a potentially important source of bias in VA models.

Table 2 presents additional estimates in the North Carolina data. Here, I replace the dependent variable from Table 1, the change in mean end-of-grade scores, with the change in mean prior-year scores for the same students. That is, when examining the change in 5th grade teachers' predicted VA between years $t-1$ and t, the dependent variable is constructed from the average 4th grade scores of the students in the teachers' 5th grade classrooms in $t-1$ and t. Grade $g-1$ scores are strongly predictive of grade-g scores, at both the individual and school-grade-year levels, but can't be causally attributed to the quality of grade-g teachers. Thus, the change in mean prior-year scores at the school-grade-cohort level should not be affected by the teacher switches that generate the variation in the independent variable, and can be used to diagnose non-randomness in the latter.

Column 1 uses the same specification as in Table 1, Column 2. The coefficient is +0.134 and is highly significant. Evidently, student quality changes

importantly when teaching staffs switch, in ways that are correlated with the change in average predicted VA that is the basis for the CFR-I quasi-experiment. The relationship is shown as a binned scatter plot in Figure 1, Panel B.[20]

After a preliminary version of this paper was shared with Chetty, Friedman, and Rockoff, they confirmed that ΔQ_{sgt} is correlated with the change in mean lagged scores in the CFR-I sample (and in a separate sample from Los Angeles as well). In a specification like that in Table 2, Column 1, albeit with year fixed effects rather than school-year effects, they obtain a coefficient of 0.226 (standard error 0.033). When I use an identical specification in the North Carolina sample, the coefficient is 0.220 (0.021).

Columns 4-6 of Table 1 suggest that sample construction may be a contributing factor. As noted earlier, CFR's estimates exclude teachers with missing predicted VA scores from the independent variable and exclude their students from the dependent variable. There is reason to expect that the sample selection created by the latter exclusion induces a positive correlation between average prior-year scores of the remaining students and the change in average predicted VA of the remaining teachers.

To see this, consider a school-grade cell with two veteran teachers in $t-1$, labeled A and B. Suppose that teacher B leaves after the year and is replaced in t by teacher C, who herself remains for only one year. Teacher C is excluded from CFR-I's calculation of Q_{sgt}, which therefore equals the predicted VA of teacher A. Because information is available for both teachers in $t-1$, $Q_{sg,t-1}$ equals the average predicted VA of A and B. Thus,

$$\Delta Q_{sgt} = \hat{\mu}_{At} - \frac{1}{2}\left(\hat{\mu}_{A,t-1} + \hat{\mu}_{B,t-1}\right). \tag{6}$$

Approximating $\hat{\mu}_{jt} = \hat{\mu}_{j,t-1} = \hat{\mu}_j$,[21] this reduces to $\Delta Q_{sgt} = \frac{1}{2}\left(\hat{\mu}_A - \hat{\mu}_B\right)$. This

[20] The first and last points appear to drive the upward slope in this Figure. But this is an illusion – the slope is similar (and remains highly significant) when the extreme changes are excluded.

[21] $\hat{\mu}_{jt}$ and $\hat{\mu}_{j,t-1}$ are constructed from the same data, the average residual scores of teacher j's students in years other than t and $t-1$. The best-prediction weights for year t differ from those for $t-1$, however, placing more weight on $t+1$ residuals and less on $t-2$ residuals.

is greater than zero if teacher A's predicted VA is higher than that of teacher B, and less than zero if B is predicted to be better than A.

Consider the former case, where $\hat{\mu}_A > \hat{\mu}_B$ and $\Delta Q_{sgt} > 0$. We can draw the following probabilistic inferences:

- Teacher A is likely an above average teacher. That is, $E\left[\mu_A \,|\, \hat{\mu}_A > \hat{\mu}_B\right] > E\left[\mu\right]$.
- Teacher C is likely worse than teacher A: $E\left[\mu_C \,|\, \hat{\mu}_A > \hat{\mu}_B\right] = E\left[\mu\right] < E\left[\mu_A \,|\, \hat{\mu}_A > \hat{\mu}_B\right]$.
- Students with higher prior-year test scores tend to be assigned to teachers with higher predicted VA (see Appendix Table 2), in both CFR's district and in North Carolina. Thus, teacher C likely is assigned students with lower prior-year scores than teacher A: $E\left[A^*_{i,t-1} \,|\, j\left(i,t\right) = C,\, \hat{\mu}_A > \hat{\mu}_B\right] < E\left[A^*_{i,t-1} \,|\, j\left(i,t\right) = A,\, \hat{\mu}_A > \hat{\mu}_B\right]$.
- The exclusion of teacher C's students from the calculation of mean prior-year scores in year t biases this upward relative to the whole-school average, with a similar effect on the change from $t-1$ to t: $E\left[A^*_{i,t-1} \,|\, j\left(i,t\right) = A,\, \hat{\mu}_A > \hat{\mu}_B\right] > E\left[A^*_{i,t-1} \,|\, \hat{\mu}_A > \hat{\mu}_B\right]$.

Each of these inferences is only probabilistic, but they all hold on average. A parallel argument implies that when $\Delta Q_{sgt} < 0$, the change in mean prior-year scores is biased downward. This implies that sample selection creates a positive correlation between the measured values of ΔQ_{sgt} and mean prior-year scores when students whose teachers are missing VA scores are excluded.

To examine this empirically, Columns 2-4 of Table 2 vary the sample used for construction of the dependent variable (the change in mean prior-year scores) and the independent variable (the change in mean predicted VA of the school-grade's teachers). In Column 2, teachers observed only in years $t-1$ and/or t are assigned predicted VA of zero and included in the independent variable. The coefficient here is similar to that in Column 1.

The approximation is thus inexact. But the difference is very small.

In Column 3, these teachers are excluded from the independent variable, but their students are included in the dependent variable. Here, the coefficient is much reduced. This is consistent with the above intuition that sample selection in the group of students included in the dependent variable creates a positive bias in the quasi-experimental regressions. However, it remains significantly different from zero. The coefficient grows again in Column 4, where all classrooms are included in both dependent and independent variables.[22] The contrast between Columns 3 and 4 is inconsistent with CFR-I's argument that the inclusion of teachers with missing predicted VA scores attenuates the coefficient. Rather, the bias associated with the exclusion from the *dependent* variable of students whose teachers do not have predicted VA scores appears to be a much more important factor. However, the significance of the Column 4 coefficient indicates that sample selection does not fully account for the evident association of the change in predicted VA with the change in students' lagged scores.

Given the evidence that the change in the average measured VA of grade-g teachers is correlated with the change in students' prior-grade scores, it is natural to wonder about dynamics in prior years, both in earlier grades for the same students and in earlier cohorts. Unfortunately, it is challenging to measure these dynamics, as the "treatment" is not observed directly but inferred from student outcomes in prior years. In many cases the same students' scores will be used for both dependent and independent variables, potentially creating spurious correlations. Nevertheless, I have been able to conduct some exploratory analyses using VA measures constructed only from years outside of the longer $\{t-3, t-2, t-1, t\}$ window. These indicate that the cohort-over-cohort change in average grade-g teacher VA is correlated with the change in cohort scores in grades $g-1$, $g-2$, and $g-3$. The evidence regarding between cohort, within school (or within school-grade-subject) trends in student

[22] For comparability to earlier columns, the Column 4 sample is restricted to school-grade-year cells where at least one teacher has non-missing predicted VA. The comparison between Columns 3 and 4, where the only difference is whether teachers with missing predicted VA scores are included in the independent variable, is inconsistent with CFR-I's argument that the inclusion of these teachers attenuates the coefficient.

performance is less clear – while the $t-1$ cohort's achievement history is correlated with ΔQ_{sgt}, there is little relationship between ΔQ_{sgt} and student achievement changes in prior cohorts. The apparent association between the change in teacher VA and the $t-1$ to t change in prior-grade achievement evidently reflects idiosyncratic cohort-level shocks rather than ongoing school-by-grade-level trends.

5 Quasi-Experimental Estimates Under A Selection on Observables Assumption

The observed association between teacher switching and changes in student preparedness limits the credibility of quasi-experimental estimates. The situation is similar to a randomized experiment, where the supposedly randomized treatment is found to be correlated with subjects' pre-determined characteristics. In such a situation, the most defensible estimator controls for the pre-determined characteristics that are correlated with treatment.[23] If selection into treatment depends only on the observed characteristics, the estimated treatment effect can be unbiased, though if this unverifiable assumption does not hold then there may be bias of unknown sign and magnitude.

I begin my exploration of selection-on-observables specifications with a graphical analysis. Figure 1, Panel C shows a binned scatter plot of the difference-in-difference in average student scores – the change from the $t-1$ cohort to the t cohort in the growth in mean scores between grades $g-1$ and g – against the predicted change due to changes in the teaching staff in the school-grade cell. As in Panels A and B, the scatterplot is quite linear. But the slope is 0.847, with a standard error (estimated from the micro-data) of 0.014. This is statistically and substantively less than one, indicating that the

[23] In personal communication, CFR suggest that the coefficients in Table 2 could reflect fluctuations in students' grade-$g-1$ scores that do not persist into grade g, implying that controls for observables are not necessary. But they do not offer any empirical evidence for this. As in all non-experimental analyses, estimates without controls *could* be unbiased even when units are non-randomly selected into treatment, but this is in general unlikely.

VA-based measures over-predict the change in student achievement growth and thus that the VA scores are biased by student sorting.

By focusing on growth scores, this plot assumes that end-of-grade scores rise one-for-one with lagged scores. Regression specifications that loosen this constraint are presented in Table 3. Panel A reports results from the North Carolina sample, while Panel B contains results from identical specifications reported by CFR-I, where available. In Columns 1-4, I follow CFR-I's preferred strategy of excluding classrooms with missing predicted teacher VA, while Columns 5-8 include these classrooms with the teacher's predicted VA set to zero.

Columns 1, 5, and 7 present estimates without controls for changes in student observables. The first two of these repeat the specifications from Table 1, Columns 1 and 4, respectively. Column 5, following CFR-I, uses year fixed effects; Column 7 repeats the specification with the school-by-year effects used in Column 1. As before, the coefficient is close to one in Column 1, but notably smaller than one in Columns 5 and 7.

CFR-I present only one specification that controls for observables. This is reported in Column 2. It includes cubic polynomials in the change in the mean prior score and mean prior other-subject score, as well as leads and lags of ΔQ_{sgt}. Results are quite similar to those in Column 1. The lead and lag terms may be endogenous, however. CFR-I construct ΔQ_{sgt} based only on data from outside the $\{t-1, t\}$ window used to compute the dependent variable, to avoid mechanical correlations between teacher VA and student outcomes. But the lead, $\Delta Q_{sg,t+1}$, is based in part on data from $t-1$ and the lag, $\Delta Q_{sg,t-1}$, is based in part on data from t.[24]

Column 3 presents an estimate that includes the polynomials in mean prior scores but excludes the lead and lag of ΔQ. The key coefficient is much reduced here, to 0.895.[25] Column 4 further restricts the controls, replacing

[24]The coefficient on the lead term, $\Delta Q_{sg,t+1}$, is 0.269 (standard error 0.016). Taken literally, this is a failed falsification test, as teachers who arrive in $t+1$ should not have any effect on scores in t. But the mechanical correlation deriving from the use of overlapping data to construct $\Delta Q_{sg,t+1}$ and $\Delta \bar{A}^*_{sgt}$ counsels against taking this failure, or the specification as a whole, too seriously.

[25]The Column 3 estimate is essentially unchanged when the sample is restricted to that

the cubic polynomials in the same-subject and other-subject prior scores with a single linear control for the same-subject mean prior score. This has little effect relative to the richer specification. I thus restrict attention in all further analyses to specifications with linear controls.

In Columns 5-8, I include all classrooms in the means used to construct the dependent and independent variables, assigning predicted VAs of zero to teachers without enough data to construct more informed predictions. As noted earlier, even without controls this approach yields a coefficient of 0.866 (in the North Carolina sample) or 0.877 (in CFR-I's sample) when year fixed effects are controlled, and coefficient of 0.818 (in the North Carolina sample) when school-year effects are included. All are significantly different from zero. Columns 6 and 8 show that both of the North Carolina estimates fall when mean predicted scores are controlled, though the decline is smaller than when classrooms with missing predicted VA scores are excluded.

CFR, responding to an earlier draft of this paper, suggested that the association between the change in VA and the change in prior-year scores could be mechanical, deriving from the use of overlapping data to estimate the two variables (CFR 2014c). Because the prior-year scores of $t-1$ students were obtained in $t-2$, and $t-2$ data is used to predict teachers' VA, shocks to $t-2$ outcomes could potentially induce a spurious association between ΔQ_{sgt} and the $t-1$ to t change in students' prior-year scores.

The most straightforward way to address this is to compute VA predictions for $t-1$ and t that exclude data from $t-2$ as well as from $t-1$ and t. There is no overlap in the data used for these "leave three out" predictions and that used to measure the change in students' prior-year scores, and thus no mechanical association. I present results from this specification in Appendix B. I also explore there other specifications suggested by CFR (2014c) to remove mechanical effects. Results are quite robust. The leave-three-out estimates indicate, if anything, a stronger association between ΔQ_{sgt} and the change in lagged outcomes than in Table 2, and a smaller ΔQ_{sgt} coefficient in the key prediction bias specification, controlling for lagged outcomes, than in Table

in Column 2 (which excludes observations for which the lead or lag of ΔQ is not available).

3. Other specifications yield small differences in results, and in a few the estimated "effect" of ΔQ_{sgt} on the change in students' lagged scores becomes statistically insignificant. But the quasi-experimental estimates of the effect on $\Delta \bar{A}^*_{sgt}$, controlling for the change in lagged scores, are always significantly below one.

Across all the specifications I have estimated, results are quite consistent: In any specification that attempts to control for changes in student observables – particularly those induced by sample selection in the construction of school-grade averages – the key quasi-experimental coefficient is significantly lower than one. Estimates are generally near 0.9 when classrooms with missing teacher VA predictions are excluded and around 0.8, or even a bit smaller, when they are included. My preferred specification is in Table 3, Column 8. This includes all classrooms, controls for lagged achievement, and identifies the effect only within school-by-year cells; it yields an estimate of the key coefficient of 0.800 (standard error 0.021). As I discuss below, in Section 7, this indicates a substantively important amount of bias.

6 Long-Run Effects

The evidence presented above that the results of CFR-I's analysis of student end-of-grade test scores are sensitive to the inclusion of controls suggests that further investigation is warranted of CFR-II's analysis of the effects of teacher VA on students' longer-run outcomes such as college graduation or earnings.

CFR-II present two sets of analyses of longer-run outcomes. The first set, and the ones that they compute for the most outcomes, are "cross-class comparisons," simple regressions of class-level mean long run outcomes on the teacher's predicted VA with controls. The second estimates, presented for a few outcomes, are quasi-experimental analyses akin to those explored above. I reproduce both. I begin in Subsection 6.1 with a discussion of the identification problem and the implications of non-random sorting of students to teachers. I then present, in Subsection 6.2, estimates of the long-run effects of North Carolina teachers, focusing on the sensitivity to the selection of controls and

to the estimation strategy.

6.1 Methods

A very simple model for students' long-run outcomes is:

$$Y_i = \sum_t \tau_{j(i,t)t} + \eta_i, \tag{7}$$

where Y_i is the outcome (e.g., high school graduation) for student i; $j(i,t)$ is the identity of the student's teacher in year t; τ_{jt} is the causal effect of teacher j on the long-run outcomes of her year-t students, holding constant other teachers' effects; and η_i is a residual that includes all non-school influences on the outcome.

CFR-II argue that direct estimates of the τs using this specification are biased, because classroom assignments are dynamic. They focus on specifications that consider the impact of one teacher at a time. (7) can be rewritten as:

$$Y_i = \tau_{j(i,t)t} + \tilde{\eta}_{it}, \tag{8}$$

where

$$\tilde{\eta}_{it} = \eta_i + \sum_{s \neq t} \tau_{j(i,s)s}.$$

(8) suggests a value-added model for teachers' effects on students' long-run outcomes, using observable characteristics measured at the end of year $t-1$ to absorb associations between the year-t teacher assignment and the elements of $\tilde{\eta}_{it}$. But CFR-II (Appendix A) argue that even this is not possible – that $\tilde{\eta}_{it}$ differs systematically across teachers, even after controlling for observables measured at the end of year $t-1$. They suggest that this is attributable to unobserved, permanent family characteristics (e.g., family connections) that are important determinants of Y_i and that are non-randomly distributed across teachers. Such sorting would seem to preclude estimation of teachers' effects on test scores as well, but CFR-II speculate that these family characteristics might not be correlated with test scores conditional on lagged scores, permitting the

estimation of test score VA.

As an alternative to estimating earnings VA, CFR-II focus on estimating the coefficient of the infeasible regression of teachers' long-run impacts on their test-score impacts:

$$\phi \equiv \frac{cov\left(\mu_{jt}, \tau_{jt}\right)}{var\left(\mu_{jt}\right)}. \tag{9}$$

Substituting in to (8), one obtains:

$$Y_i = \mu_{j(i,t)t}\phi + \check{\eta}_{it}, \tag{10}$$

where $\check{\eta}_{it} = \tilde{\eta}_{it} + \left(\tau_{j(i,t)t} - \mu_{j(i,t)t}\phi\right)$ and the latter term is by construction orthogonal to $\mu_{j(i,t)t}$. Three challenges arise in using a specification like (10) to estimate ϕ:

First, $\tilde{\eta}_{it}$ includes the effects of the student's *other* teachers, in years $s \neq t$. Insofar as $\mu_{j(i,t),t}$ is predictive of $\tau_{j(i,s)s}$, where $j\left(i,s\right)$ represents the teacher assigned to student i year $s \neq t$, this will load into the estimated ϕ coefficient. CFR-II re-define the coefficient of interest to include the projection of subsequent teachers' earnings effects onto the year-t teacher's test score VA, labeling the reduced-form coefficient κ. [26]

Second, other components of the error term are likely correlated with μ_{jt}. CFR-II find that when students' predetermined characteristics (including variables measured from tax data, such as their parents' income) are used to predict the students' long-run outcomes Y, the resulting predictions vary systematically across classrooms, even after controlling for the X variables used in the test score VA model. There is no reason to think that this variation is orthogonal to μ_{jt}, particularly given evidence (see Appendix Table 2) that μ_{jt} is correlated with students' predetermined characteristics. Even a rich set of observables may not be sufficient to absorb the correlation between $\check{\eta}_i$ and μ_{jt}, particularly given the evidence in Section 5 of this paper that $\hat{\mu}_{jt}$ is biased by student sorting on dimensions not captured by the VA model controls. Sensi-

[26] See equations 21-23 in CFR-II, online appendix A. CFR estimate that the correlation between $\mu_{j(i,t)t}$ and $\mu_{j(i,s)s}$, $s > t$, is relatively small. Insofar as the correlation between $\mu_{j(i,t)t}$ and $\tau_{j(i,s)s} - E\left[\tau_{j(i,s)s} \,|\, \mu_{j(i,s)s}\right]$ is small as well, κ is not much different from ϕ.

tivity of the κ estimates to the specific choice of controls would make it difficult to be confident in a causal interpretation of even a specification with maximal controls. For this reason, CFR-II emphasize quasi-experimental estimates of κ. Of course, the evidence above suggests that even in the quasi-experimental design it is important to control for observables.

The third challenge to overcome in estimating ϕ (or κ) is that μ_{jt} is not observed directly; only noisy estimates are available.[27] To fix ideas, suppose $\hat{\mu}_{jt}$ is free of forecast bias from student sorting and that the parameter of interest is $\tilde{\kappa}$, the μ_{jt} coefficient from an OLS regression of Y_i on μ_{jt} and Z_{it}:[28]

$$Y_i = \mu_{jt}\tilde{\kappa} + Z_{it}\tilde{\gamma} + u_{it}. \tag{11}$$

Standard errors-in-variables results imply that if an unbiased but noisy estimate of μ_{jt} is simply substituted into (11), the $\tilde{\kappa}$ coefficient will be attenuated. Recall, however, that $\hat{\mu}_{jt}$ is an Empirical Bayes estimate – an unbiased *predictor* of μ_{jt}, not an unbiased *estimate*. But the EB prediction is only unconditionally unbiased; $\hat{\mu}_{jt}$ is not an unbiased predictor conditional on Z. Thus, while the EB shrinkage factor offsets attenuation due to measurement error when the EB estimate is used on the right-hand side of a regression bivariate regression, when controls are added the estimate of $\tilde{\kappa}$ will in general be attenuated relative to what would obtain were the true μ_{jt} included as a control.

CFR-II adopt a two-step estimator of $\tilde{\kappa}$. They first regress Y_i on Z_{it} with teacher fixed effects but without controls for variation in μ_{jt} within teachers over time, then in a second stage they regress the first-stage residuals on $\hat{\mu}_{jt}$. Because the second-stage regression does not include controls, the EB shrinkage factor ensures that the $\hat{\tilde{\kappa}}$ coefficient is not biased by measurement error in $\hat{\mu}_{jt}$.

[27] I am grateful to CFR for clarifying this issue for me, in personal communication.

[28] I use the $\tilde{\kappa}$ notation to emphasize that this may not be the same as the causal coefficient κ defined above if μ_{jt} is correlated with unobserved determinants of Y_i conditional on Z_{it}; for the moment, I am concerned only with the challenge recovering the projection coefficient that would be obtained were μ_{jt} measured without error.

But this two-step estimator only identifies the projection coefficient $\tilde{\kappa}$ if the first-stage Z coefficient is a consistent estimate of $\tilde{\gamma}$. This requires restrictions on the data generating process. In particular, $\tilde{\kappa}$ would be identified if:

- $Z_{it} - \bar{Z}_j$ is uncorrelated with $\mu_{jt} - \bar{\mu}_j$, where $\bar{Z}_j$ and $\bar{\mu}_j$ are the teacher-level means (across students and years) of Z_{it} and μ_{jt}, respectively.
- The within-teacher regression of $Y_i - \mu_{jt}\tilde{\kappa}$ on Z_{it} is identical (in probability limit) to the between-teacher regression of $\bar{Y}_j - \bar{\mu}_j\tilde{\kappa}$ on $\bar{Z}_j$.

The first of these specifies that within-teacher "drift" in value-added be unrelated to any change in (observable) student assignments. The second specifies that within-teacher and between-teacher variation in measured student characteristics Z must be equally predictive of student outcomes net of teacher quality.

This last is quite restrictive. Consider that both teachers and students are clustered within schools and school assignments are decidedly non-random. Insofar as clustering depends on unobserved factors (e.g., family income or wealth) that are imperfectly proxied by Z_{ij}, one expects the school mean of Z to be a better signal of the average long-run prospects of the school's students than is the deviation of a student's Z from her school mean for her individual prospects. If so, CFR-II's two-step estimator under-controls for the between-teacher variation in Z that is the biggest threat to identification of $\tilde{\kappa}$.

Fortunately, the two-step approach is not the only way to estimate $\tilde{\kappa}$ without bias from mismeasurement of μ_{jt}. There are at least two other options, each of which requires fewer auxiliary assumptions about the data generating process than does the two-step approach. First, one can construct an alternative Empirical Bayes predictor of μ_{jt} that is suitable for use in the multivariate regression, controlling for Z. Where CFR's EB predictor shrinks the observed performance of the teacher's students toward the grand mean $E[\mu_{jt}]$ and is an unconditionally unbiased predictor of μ_{jt}, $E[\mu_{jt} \mid \hat{\mu}_{jt}] = \hat{\mu}_{jt}$, the alternative shrinks toward $E[\mu_{jt} \mid Z_{it}]$ and is an unbiased predictor conditional on Z_{it}: $E[\mu_{jt} \mid \check{\mu}_{jt}, Z_{it}] = \check{\mu}_{jt}$. This approach is potentially complex in the presence of "drift" in VA, and I leave it to future work.

Second, and simpler, one can use methods for correcting for the influence of measurement error. I pursue this approach below. Specifically, I estimate a two-stage least squares regression of Y_i on the average residual score of the teacher's students in t, $\bar{A}_{jt}$, and Z_{it}, instrumenting for the former with CFR's EB predictor, $\hat{\mu}_{jt}$.[29] Insofar as the simple OLS regression of Y_i on $\hat{\mu}_{jt}$ and Z_{it} is biased by mismeasurement of $\hat{\mu}_{jt}$, this 2SLS estimator should undo the bias. It is thus consistent under more general conditions than the restrictive assumptions required for consistency of CFR's two-step estimator. Evidence that CFR's two-step estimates diverge from the 2SLS estimates would suggest that these assumptions are not satisfied and that the former are inconsistent.

6.2 Results

I present cross-class analyses of long-run effects in Columns 1-5 of Table 4. Controls for observables are of undisputed importance here, as students are not randomly assigned to classrooms (Rothstein, 2010) and in both CFR-I's sample and the North Carolina data students with higher prior scores tend to be assigned to teachers with higher predicted VA (see Appendix Table 2).

I present results without controls in Column 1. Teacher predicted VA is strongly positively correlated with each of the five distal outcomes I consider. A one standard deviation increase in an elementary teacher's predicted VA is associated with a 0.74 percentage point increase in her students' high school graduation; a 0.86 percentage point increase in college plans; a 3.42 percentage point increase in 4-year college plans; a 0.046 increase in high school GPA, and a 1.34 rank increase in standing within the high school class.

But these associations combine student sorting and teachers' causal effects. In Column 2, I present estimates using CFR-II's two-step method to control for classroom-level mean prior scores and other characteristics (e.g., free lunch status). Across outcomes, the coefficients are reduced by one-half to two-thirds.

Column 3 presents traditional multivariate regressions, controlling directly

[29]Recall that $\hat{\mu}_{jt}$ is computed from test score residuals from years *other than* t, so measurement error is independent of that in $\bar{A}_{jt}$.

for classroom characteristics. Estimates are about one-third smaller than in Column 2. Column 4 adds teacher-level means of each of the student-level observables. This reduces the coefficients by about one-sixth more, consistent with the idea that between-teacher variation in observables is more predictive of outcomes than is within-teacher variation.

The estimates in Columns 3 and 4 might be biased downward due to measurement error in the unconditional Empirical Bayes VA measure, while those in Column 2 might be biased upward by the failure to fully control for observables. Column 5 presents 2SLS estimates where the VA measure is used as an instrument for the mean residual test score in the class. As discussed above, this is an alternative to CFR's two-step estimator that is consistent under less restrictive assumptions. 2SLS estimates are essentially identical to those in Column 4, suggesting that measurement error in $\hat{\mu}_{jt}$ is not a major problem. Like that column, they are much smaller than the estimates in Column 2, indicating that the assumptions needed to rationalize the two-step estimator are not satisfied here and that the Column 2 estimates fail to fully control for observables.[30]

The estimates in Column 5 control only for the VA model covariates, as these are all that are available in the North Carolina data. CFR-II (Table 2, Column 2) present estimates that add controls for parents' characteristics extracted from tax returns. They find – using the non-standard two-step estimator – that effects of teachers' predicted VA on college-going fall noticably when the additional covariates are included. As I do not have access to these additional controls, the more fully controlled estimates in Columns 4 and 5 of Table 4 should be seen as an upper bound to teachers' causal effects. Moreover, it is difficult to be confident that even CFR-II's expanded control vector fully eliminates student sorting bias, so causal interpretation of the estimates is tenuous.

I explore quasi-experimental estimates of the effect of teacher predicted VA on longer-run outcomes in Columns 6 and 7 of Table 4. Column 6 presents

[30]First stage coefficients are nearly exactly one, again indicating that the EB shrinkage factors nearly perfectly offset attenuation due to measurement error in $\hat{\mu}_{jt}$.

estimates without controls, while Column 6 adds a control for the change in the school-grade-subject-year mean prior-year test score.[31]

Even without controls, in Column 6, the quasi-experimental estimate is significant only for high school graduation. Two of the other coefficients are near zero, while the other two remain substantial but not large enough to distinguish from zero.[32] When the control for student sorting is added, the high school graduation coefficient falls by about half and ceases to be significant, and all of the other coefficients also fall substantially in magnitude.

7 Conclusion

This paper has implemented CFR-I's test for bias in teacher VA scores, and CFR-II's analysis of long-run outcomes, in data from the North Carolina public schools. All of CFR-I's key reported results are successfully replicated in the North Carolina sample. In particular, CFR-I's preferred quasi-experimental test indicates no bias in measured teacher VA from within-school, between-teacher sorting of students to classrooms, just as in their sample.

I also reproduce the failure of one of CFR-I's specification checks, aimed at detecting the importance of sample selection to the results: When all teachers and students, rather than just the classrooms for which predicted VA scores can be constructed, are included in the sample, the key coefficient falls and the null hypothesis is rejected. I argue that the inclusion of all classrooms allows for a closer comparison of mean achievement between years.

Further investigation shows that teacher switching does not create a valid quasi-experiment in North Carolina, even when all classrooms are included. Teacher turnover is associated with changes in student quality, as measured by the students' prior-year scores. When changes in observed student quality are

[31] In contrast to CFR-I's preferred specifications for end-of-year scores, CFR-II's samples for analyses of long-run outcomes include teachers observed only once, with predicted VA set to zero. I follow this decision.

[32] Several of the long-run outcomes, GPA and class rank in particular, are available only for a few cohorts, limiting the number of observations that can be used in the quasi-experimental analyses of cohort-to-cohort changes.

controlled, in either the CFR-I sample of teachers with non-missing predicted VA or the fuller sample that includes imputed VA scores, the key coefficient is between 0.8 and 0.9, precisely estimated, and highly significantly different from zero. The estimates that include all teachers, which I regard as more credible, are at the bottom of this range. They imply, in CFR-I's terminology, forecast bias of about 20%.

Finally, I revisit CFR-II's estimates of the effects of teacher VA on students' long-run outcomes. CFR-II find that the estimated effects in cross-sectional regressions on observational samples are modestly sensitive to controls for student observables. I show that CFR-II's methods under-control for differences in student observables across teachers, and that more conventional methods indicate substantial sensitivity in the North Carolina sample. My quasi-experimental estimates with controls for changes in student quality indicate no statistically significant effects on end-of-high-school outcomes, and yield point estimates that are uniformly smaller (more negative) than in the specifications without controls that CFR-II report.

Unfortunately, the North Carolina data do not provide as rich information about students' family backgrounds or longer-run outcomes as are available in CFR-II's data. I thus cannot fully explore teachers' long-run effects. But my results are sufficient to re-open the question of whether high-VA elementary teachers have substantial causal effects on their students' long-run outcomes, and even more so to call into question the specific magnitudes obtained by CFR-II's methods.

None of the tests that CFR-I report – with the exception of the failed specification check discussed above – would identify the violations of the quasi-experimental research design that I diagnose here. Where I am able to estimate the specifications that they report, I obtain substantively identical results in the North Carolina sample, and indeed CFR have confirmed (in personal communication) that many of my key results obtain in their data. It thus seems likely the remainder would generalize across samples as well. At a minimum, pending further evidence there is no grounds for confidence in the unbiasedness of VA measures, in the district that CFR study or elsewhere.

It is worth considering whether the statistically significant bias detected in the North Carolina sample is substantively important. In a simple model, the quasi-experimental coefficient equals the ratio of the variance of teachers' true causal effects to the variance of the sum of the causal effects and the component of student sorting bias that is constant across a teacher's classrooms. Thus, my results imply that the variance of the permanent component of student sorting bias is between 11% and 25% of the variance of teachers' true effects. The first of these is in the middle of the range that Rothstein (2009, 2010) established as consistent with the data in simulations that used the amount of observable sorting to bound the amount of sorting on unobservables, while the second is closer to the top of the range that Rothstein (2010, Section VI) argued was plausible.[33] As my estimates are quite precise, I can rule out both the very upper end of that range, corresponding to biases that swamp the signal in VA scores, and the lower end, corresponding to essentially no bias.

This suggests that policies that use VA scores as the basis for personnel decisions may be importantly confounded by differences across teachers in the students that they teach, though the problem is not likely to be as severe as would be implied by the worst-case scenarios consistent with prior evidence. Teachers who have unusual assignments may be rewarded or punished for this under VA-based evaluations. This will limit the scope for improving teacher quality through VA-based personnel policies.

An important, and unresolved, question is whether student sorting biases will be worse in high-stakes settings. When pay or continued employment depend on a high VA score, rational teachers would hesitate to accept assignments that will predictably depress their scores. If VA-based evaluations make it harder to staff certain types of classrooms, this can depress overall educational efficiency conditional on average teacher quality, potentially offsetting any benefits obtained through increases in the latter.

[33] CFR-I's VA model is most similar to Rothstein's (2010) "VAM2." 10% prediction bias corresponds almost exactly to the estimate in Table 7, Panel B of Rothstein (2010) (i.e., to a ratio of the standard deviation of the bias to that of the true effect of 0.33), while 20% prediction bias is midway between the Panel C and Panel D estimates (with a ratio of standard deviations of 0.5).

References

Chetty, R., J. N. Friedman, and J. E. Rockoff (2011): "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," Working Paper No. 17699, National Bureau of Economic Research.

——— (2014a): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, forthcoming.

——— (2014b): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, forthcoming.

——— (2014c): "Prior Test Scores Do Not Provide Valid Placebo Tests of Teacher Switching Research Designs," Downloaded October 13, 2014 from `http://obs.rc.fas.harvard.edu/chetty/va_prior_score.pdf`.

——— (2014d): "Stata Code for Implementing Teaching-Staff Validation Technique," Downloaded July 21, 2014, from `http://obs.rc.fas.harvard.edu/chetty/cfr_analysis_code.zip`.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2006): "Teacher-student matching and the assessment of teacher effectiveness," *Journal of Human Resources*, 41(4), 778–820.

Guarino, C. M., M. M. Reckase, and J. M. Wooldridge (2012): "Can Value-Added Measures of Teacher Education Performance Be Trusted?," Working paper 18, The Education Policy Center at Michigan State University.

Kane, T. J., D. F. McCaffrey, T. Miller, and D. O. Staiger (2013): "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment," Research paper, Bill & Melinda Gates Foundation, Seattle, Washington.

Kane, T. J., and D. O. Staiger (2008): "Estimating Teacher Impacts On Student Achievement: An Experimental Evaluation," working paper 14607, National Bureau of Economic Research.

Rothstein, J. (2009): "Student Sorting And Bias In Value-Added Estimation: Selection On Observables And Unobservables," *Education Finance and Policy*, 4(4), 537–571.

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

Rothstein, J., and W. J. Mathis (2013): "Review of Two Culminating Reports from the MET Project," National Education Policy Center, Boulder, CO.

Appendix

A Comparison of CFR-I estimates and North Carolina replication

Appendix Tables 1-5 present replication estimates from North Carolina in parallel with those reported by CFR-I for their unnamed school district.

Appendix Table 1 presents CFR-I's Table 2. Autocovariances are similar in the two samples for elementary English teachers, but higher in the North Carolina sample for elementary math teachers. Similarly, in English the two samplies yield nearly identical estimates of the standard deviation of teachers' VA, net of sampling error, but in math the North Carolina sample yields an estimate about one-fifth larger than does CFR-I's sample.

Appendix Figure 1 presents the autocorrelations in the two samples (also reported in Appendix Table 1). In both samples, the autocorrelations are higher in math than in reading; they are also higher in each subject in North Carolina than in CFR-I's sample. Where CFR-I found that the autocorrelations stabilize at lags longer than 7, the North Carolina sample suggests that they continue to decline out to the end of the sample.

Appendix Table 2 presents estimates from CFR-I's Appendix Table 2. These are coefficients of regressions of student characteristics on their teachers' predicted VA. Raw regression coefficients are attenuated because the predicted VA measures are shrunken, and thus have lower variance than the teachers' true effects. CFR-I multiply their coefficients by 1.56, the average ratio of the standard deviation of true effects to the standard deviation of predicted effects. I present unadjusted coefficients in Panel B and multiply these by 1.34, the corresponding ratio in the North Carolina sample, in Panel C. Estimates are broadly similar, though there is perhaps less sorting of high-prior-achievement students to high-predicted-VA teachers in North Carolina than in CFR-I's sample. Interestingly, in North Carolina high-minority-share schools seem to have teachers with higher average VA than their low-minority-share counterparts.

Appendix Table 3 presents estimates from CFR-I's Table 3. (I do not reproduce their Column 3, as their code archive does not make clear how their dependent variable is constructed.) Results are broadly similar. In Column 2, my coefficient is significantly different from zero where theirs is not, but both are small in magnitude.

Appendix Table 4 presents estimates from CFR-I's Table 4. Many of these are presented elsewhere as well; they are included here for completeness. I do not reproduce CFR-I's Column 5, as my North Carolina sample excludes

middle school grades. But all other estimates are strikingly similar between the two samples.

Appendix Table 5 presents estimates from CFR-I's Table 5, in Panel A, and from my replication in Panel B. Estimates are quite similar, despite the higher share of teachers assigned predicted VA scores of zero in Column 2 in my sample than in CFR-I's. Panel C repeats the specifications from Panel B, adding as a covariate the between-cohort change in average prior-grade scores. All of the coefficients fall in this panel; those in Columns 1 and 4 are marginally significantly different from 1 (t=1.93 and t=1.77, respectively), while the other confidence intervals clearly exclude 1.

B Potential spurious associations between changes in VA and changes in students' lagged scores

After reviewing an earlier draft of this paper, CFR confirmed (in personal communication) that the association between ΔQ_{sgt} and the change in students' prior-year scores is of similar magnitude in their sample as in the North Carolina data. They argued, however, that it was likely to be attributable to so-called "mechanical" relationships. Specifically, they noted that data from $t-2$ is used both to predict the VA of teachers in $t-1$ and t, and thus to compute ΔQ_{sgt}, and to compute the prior-year scores of $t-1$ students. In this Appendix, I show that the effects that I find are *not* due to this sort of mechanical effect.

The most straightforward way to avoid any mechanical relationship between the independent and dependent variables is to construct VA predictions for years $t-1$ and t using data that use only data from outside the $\{t-2, t-1, t\}$ window. These "leave-three-out" VA predictions are not mechanically related to the $t-1$ to t change in prior year scores, as the latter uses only data from $t-2$ and $t-1$. Estimates using these leave-three-out VA scores are presented in Row 3 of Appendix Table 6. They are quite similar to the baseline estimates (in Row 1[1]), if anything indicating larger selection problems and smaller quasi-experimental estimates. This demonstrates that mechanical effects are not an important contributor to the results.

In their response to the earlier draft of this paper, CFR suggested two alternative specifications aimed at addressing different sources of mechanical

[1]The estimates in Rows 2-7 are clustered at the school (rather than the school-by-cohort) level, as existing Stata commands do not allow clustering at the school-cohort level in IV regressions with school-year fixed effects (as in Row 7). Row 2 is identical to Row 1 but for the change in clustering; this increases standard errors by about one-third.

effects. Either source would be addressed by the leave-three-out estimator. Nevertheless, estimates of CFR's specifications, alone and in combination, are presented in the remaining rows of Appendix Table 6.

One of these specifications is designed to address teachers who follow the same group of students in multiple years as they progress across grades. If a teacher taught in grade $g-1$ in $t-2$ and then taught the same students in grade g in $t-1$, then the $t-2$ scores of those students will contribute positively both to the average VA in grade g in $t-1$ and to the average lagged scores of grade g students in $t-1$.[2] CFR propose addressing this by instrumenting for the change in VA, ΔQ_{sgt}, with a modified measure that excludes teachers who taught $g-1$ in $t-2$ or $t-1$. They find that this reduces but does not eliminate the association between ΔQ_{sgt} and the change in prior year scores.

In North Carolina, less than 4% of teacher mobility consists of teachers following students. Not surprisingly, when I modify ΔQ_{sgt} to exclude teachers who taught grade $g-1$ in $t-2$ or $t-1$, or who taught grade $g-2$ in $t-3$ or $t-2$, the modification makes little difference. The modified version of ΔQ_{sgt} is correlated 0.95 with the original version, and the first-stage coefficient is 0.98. Estimates of my key specifications, in Row 4 of Appendix Table 6, are generally similar to the baseline model; the problematic association with the change in prior-year scores is reduced only in Column 1, where classrooms with missing VA scores are excluded, and even here it is significantly different from zero.[3] Row 5 presents results that combine the leave-three-out VA scores with the no-follower IV.[4] Results are similar.

CFR's second proposed specification augments the model with school-year-subject fixed effects. This is motivated by the possibility of shocks to student test scores at the school-year-subject level that are common across student grades, due for example to accidental alignment of the school's curriculum with the test in a particular year. A negative shock in year $t-2$ would depress the predicted VA of teachers in $t-1$, which depends in part on those teachers'

[2]This is a source of a mechanical association in the differenced specification only if the teacher leaves the school or grade in t; otherwise, her VA does not contribute to the $t-1$ to t change ΔQ_{sgt}. Note also that "following" itself is a problem for the quasi-experimental analysis. This analysis is designed to test whether VA scores accurately forecast the impact of grade-g teachers on their students' learning in grade-g; if a portion of the coefficient reflects contributions that the same teachers made to students when they were in grade-$g-1$, this would need to be controlled in order to isolate the causal effect of interest.

[3]I have also explored specifications analogous to those in Columns 3 and 4 where I instrument for the change in mean prior-year scores with a modified version that excludes students of teacher "followers." This has no effect on the results.

[4]Mechanical correlations with the leave-three-out VA scores could arise from teachers who followed the students from grade $g-2$ to g.

$t-2$ performance, and would also depress the lagged scores of $t-1$ students. It would have a smaller effect on the predicted VA of teachers in t, insofar as some of these teachers are new to the school, and no effect on the lagged scores of t students. This could contribute to a positive correlation between the change in mean predicted VA and the change in mean lagged scores.

Adding school-year-subject fixed effects, in place of the school-year effects in the specifications above and in CFR-I, halves the number of degrees of freedom. Rows 4 and 5 of Appendix Table 6 present the results, first in OLS and then using the no-followers instrument. The additional fixed effects reduce the coefficients in columns 1 and 2, and three of the four are not significantly different from zero. But they have little effect in Columns 3 and 4: When classrooms with missing VA predictions are excluded, the coefficient gets closer to one with the addition of fixed effects, though the null hypothesis of one is never rejected, but when all classrooms are included the additional fixed effects make very little difference at all.

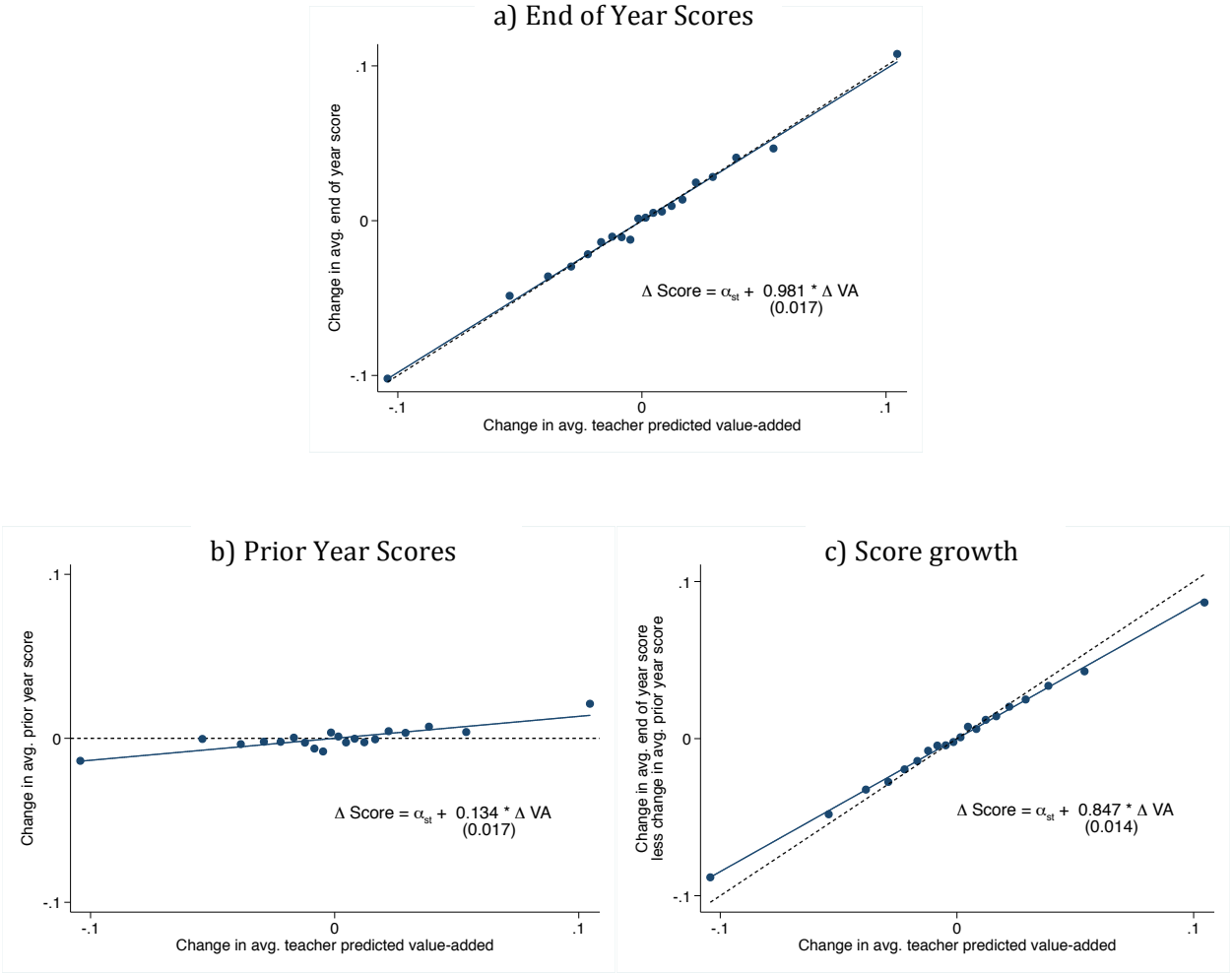
After the above results were shared with CFR, they circulated a response proposing additional specifications (CFR 2014). One excludes all data prior to $t+1$ from the VA predictions; the other instruments for the change in predicted VA with the predicted VA of the year-t teachers (i.e., for $\Delta Q_{sgt} = Q_{sgt} - Q_{sg,t-1}$ with Q_{sgt}). I have estimated each of these specifications as well. They yield results (not reported) that are broadly consistent with my baseline results: The magnitude of the association with the change in prior-year scores is similar to that in Row 1 of Appendix Table 6, and in the quasi-experimental analysis with controls for prior-year scores both indicate statistically significant forecast bias, of similar magnitude to my baseline specification.

Across a wide variety of specifications, the basic results are stable. They demonstrate that the failure of the quasi-experiment cannot be attributable to mechanical effects coming from the way that the variables are constructed. Rather, teacher mobility is not random, and the arrival of a high-VA teacher is associated with between-cohort increases in student preparedness that bias the quasi-experimental coefficient upward.

References

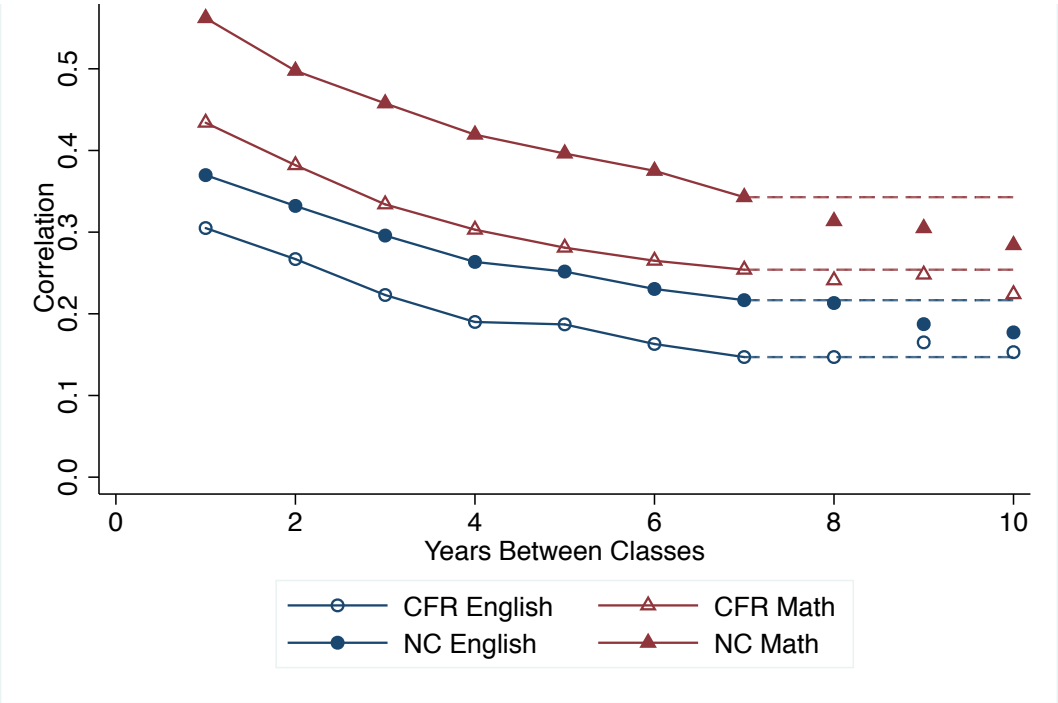
Chetty, R., J. N. Friedman, and J. E. Rockoff (2014): "Prior Test Scores Do Not Provide Valid Placebo Tests of Teacher Switching Research Designs," Downloaded October 13, 2014 from `http://obs.rc.fas.harvard.edu/chetty/va_prior_score.pdf`.

Figure 1
Effects of Teacher Turnover on End-of-Year and Prior-Year Scores



Notes: These figures are constructed using the sample used in Table 1, Column 2, pooling all grades and subjects. Each presents a binned scatter plot of cohort-to-cohort changes in school-grade-year-subject average scores against changes in school-grade-year-subject average predicted teacher VA, each residualized against school-year fixed effects. In Panel A, the score is the end-of-grade score; in Panel B, the average score for the same cohort the prior year; and in Panel C the difference between these. School-grade-year-subject cells are divided into twenty equal-sized groups (vingtiles) by the change in average predicted teacher VA; points plot means of the y- and x-variables in each group. Solid lines present best linear fits estimated on the underlying micro data using OLS with school-year fixed effects; coefficients and standard errors (clustered at the school-cohort level) are shown on each plot.

Appendix Figure 1
Replication of CFR (2014a), Figure 1
Drift in Teacher Value-Added Across Years



Notes: See notes to CFR (2014a), Figure 1.

Table 1. Replication of CFR (2014a) teacher switching quasi-experimental estimates of forecast bias

Dependent variable:	Δ Score	Δ Score	Δ Score (Predicted)	Δ Score (all students)	Δ Score	Δ Score (all students)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: North Carolina replication						
Change in mean teacher predicted VA across cohorts	1.050	0.981	0.011			0.659
	(0.023)	(0.022)	(0.011)			(0.018)
Change in mean teacher predicted VA across cohorts (with zeros)				0.866	1.203	
				(0.022)	(0.028)	
Year fixed effects	X			X	X	X
School x year fixed effects		X	X			
Grades	3 to 5	3 to 5	3 to 5	3 to 5	3 to 5	3 to 5
# of school x grade x subject x year cells	77,147	77,147	54,567	92,467	77,147	77,147
Panel B: Chetty, Friedman, and Rockoff (2014a)						
Source:	*T4C1*	*T4C2*	*T4C4*	*T5C2*		
Change in mean teacher predicted VA across cohorts	0.974	0.957	0.004			
	(0.033)	(0.034)	(0.005)			
Change in mean teacher predicted VA across cohorts (with zeros)				0.877		
				(0.026)		
Year fixed effects	X			X		
School x year fixed effects		X	X			
Grades	4 to 8	4 to 8	4 to 8	4 to 8		
# of school x grade x subject x year cells	59,770	59,770	59,323	62,209		

Notes: Panel B is taken from the indicated Tables and Columns of CFR (2014a); Panel A is estimated using the same variable construction and specifications in the North Carolina sample. Dependent variable in each column is the year-over-year change in the specified variable in the school-grade-subject-year cell. In Column 2, the dependent variable is the fitted value from a regression of end-of-year scores on parental education indicators (Panel A) or on a vector of parental characteristics taken from tax data (Panel B). In Columns 1-3 and 6, teachers observed only in years t-1 and t, whose predicted VA is set to missing by CFR's code, are excluded from the school-grade-subject-year mean predicted VA; in Columns 4 and 5, they are assigned predicted VA of zero and included. In Columns 1-3 and 5, the dependent variable is averaged only across students in classrooms whose teachers have non-missing predicted VA; in Columns 4 and 6, all students are included. See notes to CFR (2014a), Table 4 for additional details about the specifications. Standard errors are clustered by school-cohort.

Table 2. Teacher switching quasi-experimental effects on prior year scores

Dependent variable:	Δ Prior Year Score	Δ Prior Year Score	Δ Prior Year Score (all students)	Δ Prior Year Score (all students)
	(1)	(2)	(3)	(4)
Change in mean teacher predicted VA across cohorts	0.134		0.039	
	(0.021)		(0.017)	
Change in mean teacher predicted VA across cohorts (with zeros)		0.121		0.078
		(0.026)		(0.023)
School x year fixed effects	X	X	X	X

Notes: Dependent variable in each column is the year-over-year change in mean prior year scores in the school-grade-subject-year cell. In Columns 1 and 2, this is constructed only over classrooms with teachers with non-missing predicted VA scores. In Columns 3 and 4, all classrooms are included in the average. The independent variable is the change in mean predicted VA, averaged over teachers with non-missing predicted VA in Columns 1 and 3 and over all teachers (assigning zero to teachers with missing values) in Columns 2 and 4. Specifications are otherwise identical to Table 1, Column 2. Standard errors are clustered by school-cohort. Sample size in all columns is 77,177 school-grade-subject-year cells.

Table 3. Sensitivity of quasi-experimental results to controls for observables

Classrooms included in school-grade-subject means:	Classes with non-missing teacher VA				All classes, assigning 0 if VA missing			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: North Carolina								
Change in mean teacher predicted VA	0.981	0.960	0.895	0.890	0.866	0.820	0.818	0.800
	(0.022)	(0.017)	(0.015)	(0.016)	(0.022)	(0.021)	(0.022)	(0.021)
Change in mean prior year score		X (cubic)	X (cubic)	0.677		0.286		0.273
				(0.004)		(0.006)		(0.006)
Change in mean predicted end-of-year score								
School x year fixed effects	X	X	X	X				
Lead and lag changes in VA		X					X	X
Year fixed effects					X	X		
Number of cells	77,147	56,783	75,182	77,147	92,467	91,029	92,467	91,029
Panel B: Chetty, Friedman, and Rockoff (2014a)								
Source	*T4, C2*	*T4, C3*			*T5, C2*			
Change in mean teacher predicted VA	0.957	0.950			0.877			
	(0.034)	(0.023)			(0.026)			
Change in mean prior year score		X (cubic)						
School x year fixed effects	X	X						
Lead and lag changes in VA		X						
Year fixed effects					X			
Number of cells	59,770	46,577			62,209			

Notes: Columns 1 and 5 are identical to columns 1 and 4 of Table 1. Column 2 matches CFR (2014a), Table 4, Column 3. It and Column 3 include cubic polynomials in the prior-year scores in the same subject and the other subject. Column 2 also includes the lead and lag of the school-grade-subject-year change in mean predicted teacher VA; cells where this is missing are excluded. All standard errors are clustered at the school-cohort level.

Table 4. Sensitivity of effects on medium- and long-run outcomes to controls for observables

	Cross-sectional regressions					Quasi-experimental regressions	
	OLS	Two-step	OLS	OLS	2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: North Carolina sample						
Graduate high school (%)	0.74	0.38	0.24	0.21	0.21	0.36	0.20
N=1,947,791	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.18)	(0.19)
Plan college (%)	0.86	0.35	0.25	0.21	0.22	0.35	0.19
N=1,269,762	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.22)	(0.23)
Plan 4-year college (%)	3.42	1.21	0.75	0.63	0.64	-0.02	-0.35
N=1,269,740	(0.14)	(0.10)	(0.09)	(0.09)	(0.09)	(0.33)	(0.33)
GPA (4 pt. scale)	0.046	0.021	0.017	0.014	0.015	0.010	-0.003
N=945,336	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.008)	(0.008)
Class rank (100=top)	1.34	0.62	0.42	0.33	0.34	0.34	0.03
N=943,409	(0.07)	(0.06)	(0.05)	(0.05)	(0.05)	(0.24)	(0.24)
Control for student observables							
Classroom means		X	X	X	X		
Teacher means				X	X		
Change in school-grade-subject mean prior year score						X	X
	Panel B: Chetty, Friedman, and Rockoff (2014b)						
College at age 20		0.82				0.86	
N=4,170,905		(0.07)				(0.23)	

Notes: Each entry represents the estimated effect of a one standard deviation increase in predicted teacher VA from a separate regression. Row headers indicate dependent variables. Cross-sectional regressions in columns 1-5 are estimated on classroom-year-subject means. Classrooms for which the teacher is missing a leave-one-out VA score are excluded. The number of students included is listed on each row. In two-step estimates in Column 2, the dependent variable is regressed on a vector of classroom-level controls, with teacher fixed effects; the teacher effect and residual are then summed and regressed on the teacher's predicted VA without controls. In Columns 3-4 the dependent variable is regressed directly on teacher VA with covariates included as controls. Column 5 instruments for the classroom's observed mean residual score with the teacher's leave-one-out predicted VA (used as the explanatory variable in Columns 1-4). Columns 6 and 7 present quasi-experimental estimates for the annual change in mean outcomes at the school-grade-subject level. All classrooms are included in school-grade-subject-year means, assigning teachers with missing VA values of zero. Column 7 controls for the mean prior year score. Estimates in Panel B are taken from CFR (2014b), Table 2, Column 1 and Table 5, Column 1.

Appendix Table 1. Replication of CFR (2014a), Table 2
Teacher Value-Added Model Parameter Estimates

	CFR		North Carolina sample	
	Elem. School English	Elem. School Math	Elem. School English	Elem. School Math
	(1)	(2)	(3)	(4)
	Panel A: Autocovariance and Autocorrelation Vectors			
Lag 1	0.013	0.022	0.013	0.033
	(0.0003)	(0.0003)	(0.0002)	(0.0003)
	[0.305]	[0.434]	[0.370]	[0.562]
Lag 2	0.011	0.019	0.011	0.029
	(0.0003)	(0.0003)	(0.0002)	(0.0003)
	[0.267]	[0.382]	[0.332]	[0.498]
Lag 3	0.009	0.017	0.010	0.027
	(0.0003)	(0.0004)	(0.0002)	(0.0004)
	[0.223]	[0.334]	[0.296]	[0.458]
Lag 4	0.008	0.015	0.009	0.024
	(0.0004)	(0.0004)	(0.0002)	(0.0004)
	[0.190]	[0.303]	[0.264]	[0.419]
Lag 5	0.008	0.014	0.009	0.023
	(0.0004)	(0.0005)	(0.0003)	(0.0005)
	[0.187]	[0.281]	[0.252]	[0.396]
Lag 6	0.007	0.013	0.008	0.022
	(0.0004)	(0.0006)	(0.0003)	(0.0005)
	[0.163]	[0.265]	[0.230]	[0.375]
Lag 7	0.006	0.013	0.007	0.020
	(0.0005)	(0.0006)	(0.0003)	(0.0006)
	[0.147]	[0.254]	[0.217]	[0.343]
Lag 8	0.006	0.012	0.007	0.019
	(0.0006)	(0.0007)	(0.0004)	(0.0008)
	[0.147]	[0.241]	[0.213]	[0.313]
Lag 9	0.007	0.013	0.006	0.019
	(0.0007)	(0.0008)	(0.0005)	(0.0009)
	[0.165]	[0.248]	[0.187]	[0.305]
Lag 10	0.007	0.012	0.006	0.017
	(0.0008)	(0.0010)	(0.0006)	(0.0012)
	[0.153]	[0.224]	[0.177]	[0.284]
	Panel B: Within-Year Variance Components			
Total SD	0.537	0.517	0.564	0.544
Individual Level SD	0.506	0.473	0.545	0.496
Class+Teacher Level SD	0.117	0.166	0.146	0.223
Estimates of Teacher SD				
Lower Bound Based on Lag 1	0.113	0.149	0.113	0.182
Quadratic Estimate	0.124	0.163	0.121	0.194

Notes: See notes to CFR (2014a), Table 2. North Carolina estimates use CFR's code.

Appendix Table 2: Replication of CFR (2014a), Appendix Table 2
Differences in Teacher Quality Across Students and Schools

	Dependent variable: Teacher value-added						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: CFR (2014a), Appendix Table 2 (coefficients multiplied by 1.56)						
Lagged test score	0.0122			0.0123			
	(0.0006)			(0.0006)			
Special ed		-0.003					
		(0.001)					
Parent income ($10,000s)			0.00084	0.00001			
			(0.00013)	(0.00011)			
Minority (black/hispanic) student					-0.001		
					(0.001)		
School mean parent income ($10,000s)						0.0016	
						(0.0007)	
School fraction minority							0.003
							(0.003)
N	6,942,979	6,942,979	6,094,498	6,094,498	6,942,979	6,942,979	6,942,979
	Panel B: North Carolina sample (unadjusted coefficients)						
Lagged test score	0.0054						
	(0.0003)						
Special ed		0.0008					
		(0.0004)					
Minority (black/hispanic) student					0.0014		
					(0.0010)		
School fraction minority							0.016
							(0.003)
N	4,156,100	4,156,100			4,156,100		4,156,100
	Panel C: North Carolina sample (coefficients multiplied by 1.34)						
Lagged test score	0.0072						
	(0.0004)						
Special ed		0.0011					
		(0.0005)					
Minority (black/hispanic) student					0.0019		
					(0.0013)		
School fraction minority							0.021
							(0.005)

Notes: See notes to CFR (2014a), Appendix Table 2. Panel B reports coefficients from applying CFR's code to the North Carolina sample. CFR multiply their reported coefficients by 1.56 to offset the average shrinkage of the dependent variable. In Panel C, I multiply the Panel B coefficients and standard errors by 1.34, the shrinkage ratio indicated for the North Carolina sample by CFR's code.

Appendix Table 3. Replication of CFR (2014a), Table 3
Estimates of Forecast Bias Using Parent Characteristics and Lagged Scores

Dep. Var.:	Score in Year t	Pred. Score using Parent Chars.	Score in Year t	Pred. Score using Year t-2 Score
	(1)	(2)	(3)	(4)
	Panel A: CFR (2014a)			
Teacher VA	0.998	0.002	0.996	0.022
	(0.0057)	(0.0003)	(0.0057)	(0.0019)
Parent Chars. Controls			X	
Observations	6,942,979	6,942,979	6,942,979	5,096,518
	Panel B: North Carolina sample			
Teacher VA	1.011	0.010		0.022
	(0.0044)	(0.0007)		(0.0016)
Parent Chars. Controls				
Observations	4,156,100	3,571,503		2,497,994

Notes: See notes to CFR (2014a), Table 3; replication follows their methods. Dependent variables are residualized against the covariates in the VA model, at the individual level, before being regressed on on the teacher's leave-one-out predicted VA, controlling for subject. In Column 2, the second stage regression is estimated on classroom-subject-level aggregates; reported observation counts correspond to the number of student-year-subject-level observations represented in these aggregates. Standard errors are clustered at the school-cohort level.

Appendix Table 4. Replication of CFR (2014a), Table 4
Quasi-Experimental Estimates of Forecast Bias

Dependent Variable:	Δ Score	Δ Score	Δ Score	Δ Predicted Score	Δ Other Subj. Score	Δ Other Subj. Score
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: CFR (2014a)					
Change in mean teacher predicted VA	0.974	0.957	0.950	0.004	0.038	0.237
across cohorts	(0.033)	(0.034)	(0.023)	(0.005)	(0.083)	(0.028)
Year Fixed Effects	X				X	X
School x Year Fixed Effects		X	X	X		
Lagged Score Controls			X			
Lead and Lag Changes in Teacher VA			X			
Other-Subject Change in Mean Teacher VA					X	X
Grades	4 to 8	4 to 8	4 to 8	4 to 8	Middle Sch.	Elem. Sch.
No. of School x Grade x Subject x Year Cells	59,770	59,770	46,577	59,323	13,087	45,646
	Panel B: North Carolina sample					
Change in mean teacher predicted VA	1.050	0.981	0.960	0.011		0.207
across cohorts	(0.023)	(0.022)	(0.017)	(0.011)		(0.016)
Year Fixed Effects	X					X
School x Year Fixed Effects		X	X	X		
Lagged Score Controls			X			
Lead and Lag Changes in Teacher VA			X			
Other-Subject Change in Mean Teacher VA						X
Grades	3 to 5	3 to 5	3 to 5	3 to 5		3 to 5
No. of School x Grade x Subject x Year Cells	77,147	59,770	46,577	59,323		75,182

Notes: See notes to CFR (2014a), Table 4. Panel B replicates CFR's estimates using the North Carolina sample.

Appendix Table 5. Replication of CFR (2014a), Table 5
Quasi-Experimental Estimates of Forecast Bias: Robustness Checks

Specification:	Teacher Exit Only	Full Sample	<25% Imputed VA	0% Imputed VA
Dependent Variable:	Δ Score	Δ Score	Δ Score	Δ Score
	(1)	(2)	(3)	(4)
	Panel A: CFR (2014a)			
Change in mean teacher predicted VA	1.045	0.877	0.952	0.990
across cohorts	(0.107)	(0.026)	(0.032)	(0.045)
Year Fixed Effects	X	X	X	X
Number of School x Grade x Subject x Year Cells	59,770	62,209	38,958	17,859
Pct. of Observations with Non-Imputed VA	100.0	83.6	93.8	100.0
	Panel B: North Carolina sample			
Change in mean teacher predicted VA	1.140	0.866	1.044	1.033
across cohorts	(0.041)	(0.022)	(0.037)	(0.045)
Year Fixed Effects	X	X	X	X
Number of School x Grade x Subject x Year Cells	77,147	92,467	30,324	20,571
Pct. of Observations with Non-Imputed VA	100.0	68.3	94.4	100.0
	Panel C: North Carolina sample, with control for prior scores			
Change in mean teacher predicted VA	0.941	0.820	0.925	0.934
across cohorts	(0.030)	(0.021)	(0.031)	(0.037)
Change in mean prior year score	0.655	0.287	0.614	0.615
across cohorts	(0.004)	(0.006)	(0.006)	(0.008)
Year Fixed Effects	X	X	X	X
Number of School x Grade x Subject x Year Cells	77,147	91,029	30,324	20,571
Pct. of Observations with Non-Imputed VA	100.0	69.0	94.4	100.0

Notes: See notes to CFR (2014a), Table 5. Panel B replicates CFR's estimates using the North Carolina sample. In Panel C, the across-cohort change in the students' average prior-year scores is added as a control. These specifications are not reported by CFR.

Appendix Table 6. Robustness checks for quasi-experimental estimates

	Dependent variable	Δ Prior Year Score		Δ End-of-Year Score	
	Classrooms included in school-grade-subject-year means	Classes with non-missing teacher VA	All classes	Classes with non-missing teacher VA	All classes
		(1)	(2)	(3)	(4)
1	Baseline	0.134	0.078	0.890	0.800
		(0.021)	(0.023)	(0.016)	(0.021)
2	Cluster on school	0.134	0.078	0.890	0.800
		(0.028)	(0.030)	(0.021)	(0.027)
3	Using leave-three-out teacher VA predictions	0.149	0.096	0.875	0.785
		(0.031)	(0.033)	(0.023)	(0.030)
4	IV using non-following teachers	0.073	0.095	0.903	0.785
		(0.025)	(0.049)	(0.019)	(0.025)
5	Using leave-three-out teacher VA predictions; IV with non-following teachers	0.106	0.183	0.881	0.772
		(0.028)	(0.054)	(0.021)	(0.028)
6	School-year-subject FEs	0.114	*0.036*	0.924	0.813
		(0.038)	(0.040)	(0.026)	(0.034)
7	IV using non-following teachers, with school-year-subject FEs	*0.047*	*0.031*	0.940	0.799
		(0.029)	(0.050)	(0.021)	(0.028)

Notes: Specifications in Row 1 correspond to Table 2, Column 1 (Col. 1); Table 2, Column 4 (Col. 2); Table 3, Column 4 (Col. 3); and Table 3, Column 8 (Col. 4). Successive rows modify the specification. In Rows 2-7, standard errors are clustered at the school level. In Rows 3 and 4, teacher VA predictions are constructed only from data before t-2 or after t. In Rows 4, 5, and 7, the change in mean predicted teacher VA in the school-grade-subject-year cell is instrumented with a variable constructed similarly but with "looping" teachers' predicted VA set to zero. Italicized coefficients are not significantly different from zero.