Revisiting the Impacts of Teachers

Jesse Rothstein*

March 2016

Abstract

Chetty, Friedman, and Rockoff (2014a, 2014b) study value-added (VA) measures of teacher effectiveness. CFR (2014a) exploits teacher switching as a quasi-experiment, concluding that student sorting creates negligible bias in VA scores. CFR (2014b) finds VA scores are useful proxies for teachers' effects on students' long-run outcomes. I successfully reproduce each in North Carolina data. But I find that the quasi-experiment is invalid, as teacher switching is correlated with changes in student preparedness. Adjusting for this, I find moderate bias in VA scores, perhaps 10-35% as large, in variance terms, as teachers' causal effects. Long-run results are sensitive to controls and cannot support strong conclusions.

*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 three referees and conference and seminar participants at Berkeley, Northwestern, RAND, Santa Cruz, the University of Texas, the University of Wisconsin Institute for Research on Poverty, NBER, and SOLE for comments. I also thank 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 conversations.

This paper revisits the analysis and conclusions of a pair of recent papers in the *American Economic Review* that use data from New York City school records and tax filings to examine central questions about value-added (hereafter, VA) models of teacher effectiveness.[1]

The first paper (Chetty et al., 2014a; hereafter, CFR-I) attempts to measure bias in VA scores, interpreted as estimates of teachers' casual effects. Teachers' VA scores may be biased if the observed student characteristics included as controls – most notably prior scores – fail to fully absorb the unmeasured determinants of student-teacher matches, which often depend on parent requests or teacher specializations (Rothstein, 2010). CFR-I exploits teacher switches – events where one teacher exits or enters a school or grade – as plausibly exogenous changes in the quality of teachers to which students are exposed, and concludes that any biases are minimal.

The second paper (Chetty et al., 2014b; hereafter CFR-II) investigates whether a teacher's VA score is a useful proxy for her effect on longer-run outcomes, including high school graduation, college enrollment, and adult earnings. CFR-II concludes that high-VA teachers have dramatically better effects on all of these outcomes, suggesting that replacing a low VA teacher with an otherwise similar teacher with a higher VA score would bring substantial benefits for students' long-run success.

I revisit these questions in data from North Carolina.[2] Using CFR's methods and drawing on their programs (CFR 2014f), I successfully reproduce all of the key results of each paper. Further investigation, however, indicates that neither North Carolina nor New York data support CFR's substantive conclusions regarding VA bias or teachers' long-run effects.

I focus on CFR-I, as CFR-II relies on its conclusion that VA scores are unbiased. Figure 1, Panel A reproduces CFR-I's Figure 4A, which illustrates CFR-I's key result. It is a "binned scatterplot" of the cohort-over-cohort change in mean student test scores at the school-grade-subject level (on the vertical

[1]The district is unnamed in the papers. One of the authors, Raj Chetty, confirmed the district's identity in his expert testimony in the *Vergara v. California* trial.

[2]Other responses to CFR-I and CFR-II include Ballou (2012) and Adler (2013).

axis) against the change in mean predicted VA of the teachers in the school-grade-subject cell (on the horizontal axis), after residualizing each against school-year indicators. CFR-I estimate "forecast bias" (which I define more carefully below) as one minus the slope of this relationship. In the New York data, the estimated slope is 0.957 and the standard error is 0.034. Forecast unbiasedness cannot be rejected. Panel B shows the same figure as estimated from the North Carolina sample. The picture is quite similar, with a slope of 1.030 (S.E. 0.021). Given the substantial differences between New York City and North Carolina, the close correspondence is remarkable. Other results are also successfully reproduced.

When I investigate further, however, I find that teacher switching does not create a valid quasi-experiment. The treatment – the change in the average VA of the teaching staff in a school-grade cell from one year to the next – is not as good as randomly assigned but rather is correlated with pre-determined student characteristics that are predictive of outcomes. Figure 2 illustrates this. It is identical to Figure 1B, except that the vertical axis now plots the change in students' mean scores in the year *prior* to encountering the teachers whose VA scores are used to construct the horizontal axis. If the change in teacher VA were randomly assigned, the slope here should be zero. But in fact the slope is 0.144, with a standard error of 0.021.[3]

While the slope in Figure 2 is much smaller than in Figure 1B, it is significantly and substantively greater than zero. CFR (2015a) have confirmed this result in the New York data, as have Bacher-Hicks et al. (2014) in Los Angeles. Moreover, the result is not specific to test scores – I also reject a zero slope when I use on the vertical axis predictions of students' end-of-year scores based only on non-test, demographic characteristics of students such as free lunch status, race, and ethnicity (see Table 2, below).[4]

The association between VA changes and changes in student preparedness across cohorts may bias quasi-experimental estimates like those in Figure 1

[3]If the apparently influential first and last points are excluded, the slope is 0.116 (0.035).

[4]This result disproves CFR's (2015a) and Bacher-Hicks et al.'s (2014) speculation that the placebo test violation in Figure 2 is due to "mechanical" factors related to the use of test scores in constructing VA scores. See Section 3.2 and Appendix B.

relative to the causal effect of improving teacher VA, understating forecast bias. When I modify the quasi-experimental analysis to control for changes in student preparedness, the key coefficient declines notably and becomes statistically distinguishable from one. Figure 3 replaces the end-of-year scores used to measure student outcomes in Figure 1 with the change in students' scores from the end of the prior grade. These gain scores difference away factors that are beyond the current-year teacher's control, so better capture learning – and the teacher's contribution – than do unadjusted end-of-year scores. The slope in Figure 3 is 0.889 (0.015), significantly and substantively less than one. This is quite robust – across a variety of specifications that control for observed changes in student preparedness in various ways, the key coefficient is never higher than 0.93, and the confidence interval always excludes 1.

Further exploration shows that the association shown in Figure 2 is not primarily due to true endogeneity of teacher switching (as would occur, for example, if schools in gentrifying neighborhoods attract higher-VA recruits than those in declining neighborhoods), but rather is mostly an artifact of CFR-I's sample construction, which excludes a non-random subset of classrooms. When I reconstruct the analysis using all classrooms, following one of CFR-I's robustness checks, the placebo test coefficients are smaller and less robust, and the estimated slope of end-of-year scores with respect to changes in VA is both lower (0.904 in the Figure 1 specification) and less sensitive to the inclusion of controls for student preparedness.[5]

Rothstein's (2009) simulations suggested that plausible hypotheses about the amount of endogeneity in teacher VA scores imply that the prediction coefficient estimated by CFR-I should be between 0.6 and 1. My preferred estimates are around 0.85, very much in the middle of that range. Thus, rather than ruling out forecast bias in teachers' VA scores, the CFR-I quasi-experiment demonstrates that forecast bias is non-zero – not as large as might

[5]The inclusion of all classrooms requires imputing expected VA scores to teachers who lack them. My imputations follow those used by CFR-I and CFR-II. Both excluding classrooms and including them with imputed VA scores require untestable assumptions, discussed below. Appendix B explores robustness to alternative imputations, resting on different assumptions.

have been feared, but nevertheless potentially important.

The relationship between forecast bias and the magnitude of the actual biases in teachers' VA scores (which CFR-I call "teacher-level bias") depends on an auxiliary parameter – the correlation between teachers' causal effects and the bias in their scores – that is not identified by the quasi-experiment. If this correlation is assumed to be zero, as in nearly all past work, my results imply that the bias component of VA scores is 10-20% as large, in variance terms, as the component reflecting teachers' causal effects. The assumption of zero correlation is unfounded, however. If it is loosened, teacher-level bias could be as small as 4% or as large as 100% of the variance of teachers' true effects. Horvath (2015) estimates the correlation to be -0.3; if so, my estimates imply that the variance of the bias is nearly 35% of the variance of teachers' causal effects.

Bias of this magnitude would lead to substantial misclassification of teachers with unusual assignments (e.g., those thought to be particularly effective with advanced or delayed students), and thus has important implications for their use in teacher evaluations.[6] Teachers may be unfairly rewarded or punished based on the students they are assigned, and all teachers will face perverse incentives to "game" their evaluations by altering these assignments, potentially reducing allocative efficiency. Moreover, the incentives that rewards and sanctions are meant to create will be attenuated, as many will be allocated or withheld based on factors other than effective teaching.

Another implication of bias in VA scores is that inferences about the long-run effects of high VA teachers, as in CFR-II, are potentially confounded by the bias component, which is likely to be correlated with unobserved determinants of students' long-run outcomes. I turn to this in Section 4.

CFR-II present both cross-sectional and quasi-experimental estimates of the association between teachers' VA scores and their impacts on long-run earnings. I show that the cross-sectional estimates, which do not control even for observed differences in teachers' students, rely on quite restrictive assump-

[6] In Section 5, I estimate the induced misclassification rate at around 25% in a best-case scenario.

tions. Estimates that include controls, while still requiring strong (though in my view more plausible) exclusion restrictions, are more robust and, empirically, indicate much smaller (by 33-80%, depending on the outcome) long-run effects. Moreover, as in the short-run analyses of CFR-I, I find that CFR-II's quasi-experimental analyses are quite sensitive to the inclusion of controls for endogeneity of teacher switching. Indeed, none of the estimates with controls are significantly different from zero.

This comment follows an extended exchange with CFR and others (see, e.g., Rothstein, 2014; CFR 2014d; 2014e; 2015a; and Bacher-Hicks et al., 2014). The empirical results are remarkably robust across quite disparate settings. However, while productive, the exchange has not led to consensus on the interpretation of the results. I interpret them to indicate that the teacher-switching research design does not provide the credibility of a successful quasi-experiment. What evidence there is indicates that (a) VA scores are meaningfully, but not overwhelmingly, biased by student sorting, with "forecast bias" around 15% and (under reasonable assumptions) actual bias 10-35% as large, in variance terms, as teachers' causal effects, and (b) teachers' VA scores are less informative than is implied by CFR-II's results, and perhaps completely uninformative, about the teachers' long-run impacts.

1 Teacher VA, bias, and the teacher switching quasi-experiment

This section develops notation and describes CFR-I's teacher switching quasi-experimental research design and my test of it. I follow CFR-I's notation where possible; readers are referred to their paper for a more complete description.

1.1 Teacher value-added

Anecdotally, classroom assignments depend on the school's assessment of the student's ability and personality, on parental preferences (and on parents' effectiveness at getting their preferences met), on teachers' specializations, and

on factors that are idiosyncratic from the school's perspective (e.g., the date that the student enrolls). All of these may correlate with students' potential and preparedness.

The above factors are not measured, so cannot be controlled directly. VA models attempt to limit the resulting bias in estimates of teachers' causal effects on their students' end-of-year test scores by controlling for those characteristics which are observed. The most important of these factors is the student's prior test score, but some models (including CFR-I's) also control for earlier scores, free lunch status, disability, English proficiency, mobility, race, and gender. CFR-I, unusual among VA models, also include classroom- and/or school-level means of the individual controls.[7]

CFR-I's VA model has several steps. Let A_{it}^* be the test score of student i at the end of year t with teacher $j\,(i,t)$, and let X_{it} be a vector of observed covariates. First, A_{it}^* is regressed on X_{it} with teacher fixed effects:

$$A_{it}^* = \alpha_{j(i,t)} + X_{it}\beta + \epsilon_{it}. \tag{1}$$

Second, the $X_{it}\beta$ term is subtracted from A_{it}^* to form a residual score:[8]

$$A_{it} \equiv A_{it}^* - X_{it}\hat{\beta} = \hat{\alpha}_{j(i,t)} + \hat{\epsilon}_{it}. \tag{2}$$

Third, this residual score is averaged to the teacher-year level to obtain $\bar{A}_{jt}$. This is CFR's basic estimate of the effect of teacher j on her year-t students, denoted μ_{jt}. Finally, the teacher's sequence of mean residuals across other years $t' \neq t$ is used to form a leave-one-out forecast of the teacher's residual in year t, $\hat{\mu}_{jt} \equiv E\left[\bar{A}_{jt} | \left\{\bar{A}_{jt'}\right\}_{t' \neq t}\right]$. CFR-I's specific calculation of this forecast is complex and designed to accommodate the possibility that μ_{jt} may evolve ("drift") over time. For my purposes, it suffices to note that $\hat{\mu}_{jt}$ is a shrinkage

[7] The models used for actual evaluations generally use fewer controls (see, e.g., SAS Institute, 2015; American Institutes for Research, 2015; Value-Added Research Center, undated).

[8] The teacher fixed effects in (1) make little difference: In the North Carolina sample, the correlation between A_{it}, as defined in (1) and (2), and the residual from an OLS regression of A_{it}^* on X_{it} without fixed effects is over 0.99 at the student level and 0.98 at the classroom level.

estimator, which can be seen as an Empirical Bayes (EB) prediction of the teacher's causal effect μ_{jt} under the assumption that $\bar{A}_{jt}$ is a noisy but unbiased estimate of μ_{jt}.[9] Importantly, $\hat{\mu}_{jt}$ is an unbiased prediction of $\bar{A}_{jt}$ by construction, whether the latter is an unbiased estimate of μ_{jt} or not.

CFR-I refer to the EB prediction $\hat{\mu}_{jt}$ as teacher j's value-added. For clarity, I reserve that term for the true causal effect μ_{jt}, and I refer to $\hat{\mu}_{jt}$ as the *predicted* or *forecast* value-added. Hereafter, I will assume for simplicity of exposition that $\mu_{jt} \equiv \mu_j$ – that teachers' causal effects do not "drift." Empirically, however, I follow CFR-I's methods, which do not impose this.

1.2 Bias in VA estimates and predictions

The goal of VA models is not to forecast teacher residuals, but to measure a teacher's causal effect on her students. A central question in the VA literature is whether the available controls are sufficient to permit this, or whether some teachers are systematically assigned students who are unobservably advantaged or disadvantaged, conditional on the VA model controls (Rothstein, 2010, 2009; Guarino et al., 2012). In the above notation, $\bar{A}_{jt}$ may overstate μ_j for teachers whose students are systematically but unobservably stronger than expected given their Xs, and understate it for those with unobservably weaker students. If the same teachers tend to be assigned the same types of students each year, then $\hat{\mu}_{jt}$ will also be biased as a predictor of μ_j.

Consider separating the mean residual $\bar{A}_{jt}$ into four components:

$$\bar{A}_{jt} = \mu_j + b_j + v_{jt} + e_{jt}. \tag{3}$$

The first term, μ_j, represents the teacher's causal effect. The second and third terms derive from non-random student assignments that create systematic differences in ϵ_{it} across classrooms: b_j is the component that is permanent within teachers, while v_{jt} varies across years. The former might capture teacher spe-

[9] I define bias more carefully below. For the moment, the necessary assumption for $\hat{\mu}_{jt}$ to be an unbiased prediction of the causal effect μ_{jt} is that $\bar{A}_{jt} - \mu_{jt}$ is mean independent across years within teachers – that any non-randomness in student assignments in any year is not persistent across years.

cializations – a teacher who is thought to be particularly effective with, say, hyperactive students might be assigned the same students year after year – and the latter might arise if classroom groupings are non-random but classrooms are distributed randomly across teachers. I assume that v_{jt} is serially uncorrelated.[10] The final term, e_{jt}, is a noise term that is also independent across years. It includes pure sampling error and idiosyncratic classroom-level shocks such as the proverbial dog barking on test day.

The shrinkage procedure in the final step of CFR-I's model is designed to isolate the component of $\bar{A}_{jt}$ that is stable across years. In effect, this treats the idiosyncratic bias term v_{jt} as noise, comparable to e_{jt}. But the method does not isolate μ_j from b_j, which CFR-I refer to as "teacher-level bias." Thus, a central goal in the VA literature is to measure $V(b_j)$, and in particular to test whether $V(b_j) = 0$.

CFR-I define "forecast bias" as $B \equiv 1 - \lambda$, where:

$$\lambda \equiv \frac{cov(\mu_j, \hat{\mu}_{jt})}{V(\hat{\mu}_{jt})} = \frac{V(\mu_j) + cov(\mu_j, b_j)}{V(\mu_j) + V(b_j) + 2cov(\mu_j, b_j)}.$$

The second equality here follows from $\hat{\mu}_{jt}$'s construction as an Empirical Bayes prediction of $\mu_j + b_j$. Zero forecast bias ($\lambda = 1$, $B = 0$) is necessary but not sufficient for $\hat{\mu}_{jt}$ to be teacher-level unbiased (i.e., for $V(b_j) = 0$). In particular, if $cov(\mu_j, b_j) < 0$ then λ can equal or exceed one even when $V(b_j) > 0$. The available evidence suggests this is empirically relevant: Horvath (2015) estimates $corr(\mu_j, b_j) = -0.3$ for North Carolina teachers, while Angrist et al. (2015b) estimate a correlation of -0.23 (with a large standard error) between schools' causal effects and the bias in school-level VA scores in Boston.

Rothstein (2009; see also Guarino et al., 2012) attempts to quantify the magnitude of biases in common VA models, using the distribution of observables across classrooms and assessments of the likely role for unobservables. Assuming that $corr(\mu_j, b_j) = 0$, he concludes that the plausible range for λ

[10] This is restrictive – it does not allow, for example, for an autoregressive component of student assignments. I adopt the decomposition for simplicity of exposition. In practice, any non-zero covariance between $b_j + v_{jt}$ and $b_j + v_{j,t+1}$ would create bias in VA-based evaluations, which are typically based on just two or three years of data.

is roughly 0.6 to 1, corresponding to $^{V(b_j)}/_{V(\mu_j)}$ between zero and $\frac{2}{3}$. If the correlation is instead -0.3, the upper bound of the variance ratio is about 0.75.

1.3 The teacher-switching quasi-experiment

CFR-I build on an experiment conducted by Kane and Staiger (2008) in which students were randomly assigned. Let $\hat{\mu}_{jt}$ be a shrunken / Empirical Bayes prediction based on observational data from years other than t. Random assignment in t ensures that any determinants of the teacher's students' mean outcomes in that year, other than the teacher's own causal effect μ_j, are orthogonal to both b_j and $\hat{\mu}_{jt}$. Thus, a regression of these mean experimental outcomes on the observational prediction $\hat{\mu}_{jt}$ identifies λ.

Unfortunately, it has proven difficult to randomize students to classrooms at a large scale, so experimental estimates of λ have standard errors around 0.2 or higher (Kane and Staiger, 2008, Kane et al. 2013; see also Rothstein and Mathis 2013) and have not substantially narrowed the plausible range.[11]

CFR-I generalize the experimental test to a non-experimental setting, exploiting episodes where a teacher enters or leaves a school or switches grades within the school. The replacement of one teacher with another should lead to an increase in student achievement equal to the difference between the teachers' causal effects. If the teachers' VA scores are unbiased estimates of their respective causal effects, then the difference in Empirical Bayes predictions should forecast this difference without bias and scores should, on average, rise by as much as predicted. By contrast, bias in the VA scores would mean that the difference in causal effects will tend to be smaller (closer to zero) than the prediction by a factor B.

Without random assignment within schools, new and old teachers may be assigned differently selected students, reproducing the non-experimental bias in mean outcomes. To abstract from this, CFR-I aggregate to the school (s) - grade (g) - subject (m) - year (t) level and consider changes in the *average*

[11] In a very similar analysis of school-level VA scores, Angrist et al. (2015b) estimate $\hat{\lambda} = 0.86$ (S.E 0.08). They go on to develop a more powerful test of the sharper null hypothesis that $V(b_s) = 0$ and reject this. See also Deutsch (2013).

predicted VA of the teaching staff.[12] Their primary analyses regress the year-over-year change in mean student scores, $\Delta A^*_{sgmt} \equiv \bar{A}^*_{sgmt} - \bar{A}^*_{sgm,t-1}$, on the difference in mean predicted VA of the teachers to which the students were exposed (which they denote ΔQ_{sgmt}), with year or school-by-year fixed effects.[13] Their primary conclusions are based on this regression.

For aggregation to the school-grade-subject-year level to eliminate student sorting biases, it is essential that all students in the cell be included. As I discuss below, in practice CFR-I exclude a non-random subset of classrooms from their aggregates. This biases the quasi-experimental coefficient toward the observational regression of $\bar{A}_{jt}$ on $\hat{\mu}_{jt}$, which necessarily – by virtue of the Empirical Bayes shrinkage used to construct $\hat{\mu}_{jt}$ – has a coefficient of one regardless of the presence or absence of forecast or teacher-level bias.

1.4 Assessing the quasi-experiment

The regression of ΔA^*_{sgmt} on ΔQ_{sgmt} identifies λ under CFR-I's Assumption 3 (hereafter, "A3"):

> ASSUMPTION 3 (Teacher Switching as a Quasi-Experiment): Changes in teacher VA across cohorts within a school grade are orthogonal to changes in other determinants of student scores.[14]

This assumption would be violated if, for example, schools that are gentrifying

[12] For their quasi-experimental analyses, CFR-I use "leave-two-out" predictions of the year-t and $t-1$ residuals, which they denote $\hat{\mu}_{jt}^{-\{t-1,t\}}$ and $\hat{\mu}_{jt-1}^{-\{t-1,t\}}$, that are based on data from other years. I also use leave-two-out predictions, but retain the $\hat{\mu}_{jt}$ notation.

[13] CFR-I's discussion (p. 2617) suggests that the appropriate dependent variable is the change in mean *residual* scores, as defined in (2). If ΔQ_{sgmt} were randomly assigned, either raw or residual scores should yield unbiased estimates of λ. CFR-I's empirical analysis uses mean raw scores on the grounds that "changes in control variables across cohorts are uncorrelated with ΔQ_{sgmt}," (p. 2618). I show below that this is not the case.

[14] An additional assumption, unstated by CFR-I, is required to support the aggregation of Empirical Bayes predictions: Both μ_j and b_j must be independent across teachers within school-grade-subject-year cells and between outgoing and incoming teachers. The evidence suggests this assumption is counterfactual, though perhaps not by enough to matter. CFR (2015a) report that the correlation of teachers' (shrunken) VA within schools is approximately 0.2 in New York; in North Carolina, it is around 0.15. See additional discussion below and in the Appendix.

– with later cohorts more advantaged than earlier cohorts – are able to attract teachers that have higher (measured) VA than those who they are replacing.

A3 is not directly testable. But it is unlikely to hold if the change in student characteristics at the school-grade-subject-year level is correlated with ΔQ_{sgmt}. Tests like this are a standard approach to probing the validity of a quasi-experiment, and are analogous to tests commonly conducted to assess successful randomization in true experiments. The most useful characteristics for such a test are those that are predictive of outcomes but are not caused by grade-g teachers. 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.

CFR-I present a test of this form, using characteristics (household income, homeownership) that are not included in the VA specification. They interpret their null result (CFR-I, Table 4, column 4, reproduced below as column 3 of Table 1) as evidence in support of the assumption. But there is no reason not to also examine variables that *are* included in the VA model's X_{it} vector. Indeed, these characteristics are the most important to examine, as they are chosen specifically to be strong predictors of students' end-of-year scores so orthogonality failures have great potential to create bias in estimation of λ.

Below, I find that X_{it} does change across years in ways that are correlated with ΔQ_{sgmt}. I begin with prior-year scores – VA models use these to capture many otherwise hard to measure determinants of teacher assignments and of end-of-year scores – but I also obtain similar results with the full score prediction $X_{it}\hat{\beta}$ (see equation 1) and with a more restricted prediction based only on non-test elements of X_{it} (e.g., free lunch status, race, exceptionality) that are not plausibly influenced by past teachers.

The obvious explanation is that A3 is violated. The Appendix considers and rules out several potential "mechanical" explanations, proposed by CFR (2015a; 2014d) and Bacher-Hicks et al. (2014) following circulation of an initial draft of this comment, that might lead to rejections of the placebo test null even if the underlying design is valid. Further exploration indicates, however, that another mechanical explanation is an important factor. Specifically, much

of the problem derives from CFR-I's omission of teachers with missing VA predictions – those who are observed in only a single year – from their analyses. These teachers are not randomly selected, and the exclusion of their students from school-grade-subject-year averages incorporates some of the observational student-teacher sorting into the putative quasi-experiment.

This points to two alternative routes toward reducing bias in $\hat{\lambda}$ from endogeneity of ΔQ_{sgmt}. One can control for observables that are correlated with ΔQ_{sgmt}, under a selection-on-observables assumption, or one can include the missing classrooms in the school-grade-subject-year means. Each requires assumptions (as, of course, does CFR-I's strategy of excluding a non-random subset of classrooms). I pursue both options. Empirically, results are sensitive to doing *something* about the failure of the quasi-experimental research design, but mostly insensitive to just how it is addressed. In particular, results are similar across several methods for controlling for student preparedness and in specifications designed to "block" possible channels by which prior-grade scores could be an intermediate outcome of the current-grade teachers' VA. The robustness of the adjusted results raises confidence in their validity. Appendix B further explores the inclusion of missing classrooms in the sample, demonstrating that results are similarly stable when I vary the strategy for assigning VA predictions to the missing teachers and or restrict the sample to school-grade-year cells with no missing data, as suggested by CFR (2015a).

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. North Carolina is a dramatically different setting from New York City. Nearly half of North Carolina schools are rural. Education is provided by 219 separately administered districts (though the state Department of Public Instruction (DPI) plays a larger role than in many other states); New York City has a single district divided into administrative sub-districts. Just over 25% of students in North Carolina are Black

and under 15% are Hispanic, with the remainder overwhelmingly white; in New York, about 30% are Black, 40% are Hispanic, 15% are Asian, and only 15% are white non-Hispanic.

North Carolina administers end-of-grade tests in math and reading in grades 3 through 8. Third grade students are given "pre-tests" in the Fall; I treat these as grade 2 scores.[15] 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, whose 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. Instead, I assign each proctor who is not the classroom teacher a new ID that is unique to the test year.[16] This ensures that student achievement data is not used to infer the proctoring teacher's impact.

Several of CFR-I's covariates – absences, suspensions, enrollment in honors classes, and foreign birth – are unavailable in the North Carolina data. Thus, my X_{it} vector has a subset of CFR-I's controls: 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 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-

[15]Pre-test scores are missing after 2008, as well as for math in 2006 and reading in 2008. Third graders with missing pre-test scores are excluded. When students re-take the tests, I use only the score from the first administration.

[16]I use a less restrictive threshold for a valid assignment than in past work (e.g., Clotfelter et al., 2006; Rothstein, 2010). Insofar as I fail to identify non-teacher proctors, this will attenuate the within-teacher autocorrelation of $\bar{A}_{jt}$. This autocorrelation is larger in my sample than in CFR-I's. See Figure A1 in the Appendix.

grade mean prior scores; and class size.[17] For long-run outcomes, CFR-II draw on IRS data. Lacking this, I draw more proximate outcomes from high school transcripts (graduation, GPA, class rank) and exit surveys (college plans).

I start with over 8.6 million student-year-subject observations, spread across three grades (3-5), two subjects (math and reading), 1,723 schools, and 15 years (1997-2011). After excluding students with missing test scores, special education classes, and classes with fewer than 10 students, I am left with 7.1 million observations, of which 79% are linked to 36,451 valid teachers. My original sample is a bit smaller than CFR-I's, which contains approximately 18 million student-year-subject observations, but the sample size for VA calculations is similar (7.1 million vs. 7.6 million in CFR-I's sample). I have non-missing leave-one-out predicted VA scores for 257,066 teacher-year-subject cells, with an average of 22 students per cell. The sample for the quasi-experimental analysis consists of school-grade-subject-year cells with non-missing ΔQ_{sgmt}. I have 79,466 such cells, as compared with 59,770 in CFR-I.

3 The Teacher-Switching Quasi-Experiment: Reproduction and Assessment

3.1 Reproducing CFR-I's analysis in North Carolina data

I use CFR's (2014f) Stata programs to reproduce their VA calculations and analyses in the North Carolina data. Table 1 reports CFR-I's main quasi-experimental specifications (Panel A) along with corresponding estimates from the North Carolina data (Panel B). Column 1 presents coefficients from a regression of the year-over-year change in average scores at the school-grade-subject-year level (ΔA^*_{sgmt}) on the change in average predicted VA (ΔQ_{sgmt}), with year fixed effects.[18] Column 2 repeats the specification with school-year

[17] Free lunch, limited English, and special education measures are missing in some years. I set each to zero if missing, and include indicators for missing values (as well as class- and school-year means of these) in X.

[18] Following CFR-I, the regression is weighted by the number of students in the school-grade-subject-year cell; standard errors are clustered at the school-cohort level; and class-

fixed effects.

The coefficients of these regressions estimate λ under assumption A3. If this assumption holds, the null hypothesis of no forecast bias corresponds to $\lambda = 1$, while we would expect $\lambda < 1$ if teacher-level bias is present and not too negatively correlated with teachers' causal effects. My estimate in Column 1 is somewhat larger than CFR-I's, and significantly greater than 1, but when I add school-year fixed effects in Column 2, the coefficient is much smaller and, like CFR-I's, indistinguishable from the null hypothesis. This is the specification illustrated in Figure 1.

CFR-I report a placebo test of their quasi-experimental design based on changes in *predicted* scores where predictions are made using only variables that are unaffected by teacher assignments. Specifically, CFR-I regress observed scores on parent characteristics, then average the fitted values at the school-grade-subject-year level, difference across years, and use this as the dependent variable in the quasi-experimental regression. This specification is reported in Column 3 of Table 1.[19] In both samples, the year-on-year change in mean predicted VA is uncorrelated with the change in mean predicted scores.

Column 4 presents a specification drawn from CFR-I's Table 5, Column 2. In Columns 1-3, teachers who do not have leave-one-out VA predictions – because they are observed only in $t-1$ or t – are excluded from the school-grade-subject-year VA mean, and their students are excluded from the test score average. In Column 4, all teachers and students are included, with teachers with missing predictions assigned the grand mean VA score of zero. In both the New York and North Carolina samples, this leads to rejection of the null hypothesis that $\lambda = 1$, with $\hat{\lambda} = 0.88$ in New York and $\hat{\lambda} = 0.94$ in North Carolina. I discuss this result in more depth in the next subsection.

Appendix A presents reproduction estimates for most of CFR-I's other analyses. Results are generally quite similar in North Carolina as in CFR-I's

rooms with teachers not seen in other years are omitted from both dependent and independent variables.

[19] 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.

sample. I summarize the few differences briefly here. Math VA is more variable in North Carolina, while English VA has a similar variance in the two samples. In both math and English, the autocorrelation of teacher VA across years is higher in the North Carolina data (Appendix Table A2 and Appendix Figure A1), implying less noise in the measurement process and perhaps also less drift in teachers' true VA. While students with higher prior-year scores tend to be assigned to teachers with higher predicted VA in both samples (Appendix Table A7), special education students get higher VA teachers in North Carolina, on average, but lower VA teachers in New York. In North Carolina but not in New York, minority (black and Hispanic) students are assigned to teachers with lower VA, on average, but in each district the relationship between school minority share and average teacher VA is insignificantly different from zero.[20]

3.2 Assessing the Validity of the Quasi-Experiment

CFR-I's main placebo test (see Table 1, Column 3) is based on permanent parental characteristics, taken from tax returns. But these are unlikely to capture the dynamic sorting that Rothstein (2010) found to be a potentially important source of bias in VA models. Moreover, they are not observed by school administrators, so are unlikely to affect teacher assignments directly.

Panel A of Table 2 presents additional placebo test estimates in the North Carolina data. Each entry represents a separate quasi-experimental analysis, using the same specification as in Table 1, Column 2, but varying the dependent variable. In Column 1, the dependent variable is the between-cohort change in mean prior-year scores for the same students used for the quasi-experimental analysis. That is, when examining the change in the mean predicted VA of 5th grade teachers at school s between years $t-1$ and t, the dependent variable is the change in average 4th grade scores across the same two cohorts (i.e., from $t-2$ to $t-1$). Grade $g-1$ scores are strongly predictive of grade-g scores, at both the individual and school-grade-subject-year levels, so a correlation with ΔQ_{sgmt} would indicate that the quasi-experiment

[20]Bacher-Hicks et al. (2014) find that teacher VA is significantly *lower* in high minority share schools in Los Angeles.

is not valid (subject to potential caveats discussed below). The coefficient is +0.144 and is highly significant. (This is the specification illustrated in Figure 2.) Evidently, changes in student preparedness are correlated with the quasi-experimental treatment, the change in average predicted VA.

After a preliminary version of this paper was shared with CFR, they confirmed that this result holds in New York as well. In a specification like that in Table 2, Column 1, albeit with year fixed effects rather than school-year effects, CFR (2014d) report a coefficient of +0.226 (standard error 0.033). When I use an identical specification in the North Carolina sample, the coefficient is +0.231 (0.021); Bacher-Hicks et al. (2014) report a +0.268 (0.039) coefficient in data from Los Angeles.

Column 2 of Table 2 repeats the placebo test, this time using predictions of end-of-year scores based on *all* of the covariates included in the VA specification rather than just the prior-year score. That is, the dependent variable here is the cohort-over-cohort change in the mean of $X_{it}\hat{\beta}$, from equation (1). As $\Delta\bar{A}^{*}_{sgmt} = \Delta\bar{A}_{sgmt} + \Delta\bar{X}_{sgmt}\hat{\beta}$, this is scaled to correspond exactly to the bias in the quasi-experimental results deriving from the use of unadjusted scores, A^{*}_{it}, in place of adjusted scores A_{it} (see footnote 13). The coefficient is 0.105 and is again highly significant.

These results indicate that assumption A3 is violated – the change in average VA across cohorts is correlated with other determinants of the change in outcomes, so the association between the former and the latter does not identify λ. Responding to a preliminary draft of this comment, however, CFR (2014d; 2014e) suggest that the results reflect a problem with the placebo test rather than with the research design:

> Because teacher VA is estimated using data from students in the same schools in previous years, teachers will tend to have high VA estimates when their students happened to do well in prior years. Regressing changes in prior test scores on changes in teacher VA effectively puts the same data on the left- and right-hand side of the regression, mechanically yielding a positive coefficient. (CFR 2014d, p. 1)

CFR point to two potential sources of such "mechanical" effects. First, some teachers who teach grade-g students in t or $t-1$ might have taught the same cohorts of students previously, in grade $g-1$ in $t-1$ or $t-2$ (or in grade $g-2$ in $t-2$ or $t-3$). This could induce a positive correlation between the teachers' effectiveness and the students' $g-1$ scores – in effect, these prior-year scores are intermediate outcomes of the effectiveness of the grade g teacher. Second, even when teachers do not follow students across grades, a mechanical effect could arise from the fact that data from $t-2$ is used both to measure the prior-year achievement of $t-1$ students and to forecast the $t-1$ teachers' VA. Any shock that is common across grades in the school-year cell could create a positive correlation between the *measured* VA of the $t-1$ teachers and the $t-2$ scores of the $t-1$ students, biasing the placebo coefficient upward.[21]

Column 3 of Table 2 presents an alternative placebo test that excludes all mechanical effects related to test score dynamics or VA measurement by removing test scores entirely from the dependent variable. Here, I form a predicted score for each student, $X_{it}\hat{\beta}$, using the same methods as in Column 2 but using only the demographic variables – the students' age and indicators for gender, ethnicity, free lunch, special education, limited English, grade repetition, and for missing values for each of these, along with class and school-year means – in X_{it}. None of these would be affected by prior teachers' effectiveness or by school-level shocks. But I find that the change in mean predicted VA is significantly associated with the change in the mean predicted score based on these demographic characteristics alone.[22] This conclusively establishes that the placebo result cannot be attributed to the mechanical explanations proposed by CFR (2015a). [23]

So what *does* drive the placebo effect? The data point to a third mechanical

[21]Note that either dynamic would likely invalidate not just the placebo test but also CFR-I's quasi-experimental research design itself. See Appendix B.

[22]The coefficient is smaller here than in Column 2. The demographic variables are less predictive of A^*_{it} than is the full X_{it} vector. The decline in the coefficient is exactly what one would expect if ΔQ_{sgmt} is correlated both with the demographic characteristics and with prior scores conditional on demographics; see Altonji et al. (2005).

[23]Appendix B explores this issue further. While there is some evidence that "teacher followers" contribute to the effect, the results are generally quite stable.

explanation as an important factor. Recall that CFR-I's explanatory variable is constructed from predicted VA scores of teachers in $t-1$ and t, based on the residual scores of the teachers' students in years other than $t-1$ and t. If a teacher is observed in only $t-1$ or t, there is no other information on which to base the prediction. CFR-I drop the teacher from the average Q_{sgmt} and drop the teacher's students from the average $\bar{A}_{sgmt}$.

This sample selection can reintroduce student sorting into the quasi-experiment, even if teacher switching is random. In both North Carolina and New York, more advantaged students (those with higher prior scores, or with higher family income) tend to be assigned to higher VA teachers (see Appendix Table A7). So when we lack a predicted VA score for a high (respectively, low) VA teacher, excluding her from the VA average tends to reduce (increase) Q_{sgmt}, while excluding her students from the mean prior-year or end-of-year score tends to reduce (increase) $\bar{A}_{sgmt}$. This pushes both $\hat{\lambda}$ and the placebo coefficient upward relative to what would be obtained were all teachers and classrooms included.

Recall from Section 3.1 that CFR-I present one specification that includes these teachers, assigning them predicted VA scores equal to the grand mean.[24] This is not an ad hoc imputation, but rather the score implied for these teachers by the Empirical Bayes methodology. The VA prediction used in the quasi-experimental analysis is the leave-two-out prediction based on the teacher's observed performance in years other than $t-1$ and t, shrunken toward the grand mean. For a teacher observed only in those years, there is no signal at all, so shrinkage is complete and the best predictor (and the Empirical Bayes estimate) is the grand mean $\hat{\mu}_{jt} = 0$. In their Table 5, Column 2 (reproduced as Table 1, Column 4 here), CFR-I assign this grand mean to teachers observed in just a single year, and include both the teachers and their students in the school-grade-subject-year means.[25]

[24]These teachers are included as well in CFR-II's preferred quasi-experimental specifications, with a sample excluding them used only for a specification check.

[25]Teachers observed in both $t-1$ and t but no other years also have missing leave-two-out predictions. Across all their specifications, CFR-I always include these teachers, with predictions set equal to the grand mean. The issue here concerns only those teachers

I use this approach to include all classrooms in the sample in Panel B of Table 2. The placebo test coefficients are uniformly smaller here, suggesting that sample selection is an important contributor to the endogeneity identified in Panel A.[26]

The use of the grand mean for teachers missing leave-two-out VA predictions relies on an assumption that teacher VA is independent across teachers within a school. Indeed, this assumption is implicit in CFR-I's entire quasi-experimental analysis. Although CFR-I construct their predictions at the level of the individual teacher, the relevant prediction for the quasi-experimental analysis is at the level of the school-grade-year mean. If VA is not independent within schools, the average of teacher-level EB predictions is not an unbiased prediction of the average of the teachers' true effects.

In particular, if μ_j is positively correlated among teachers at the same school, the change in the average of teachers' EB predictions overstates (in magnitude) the EB prediction of the change in the average teacher's VA, even if data is available for all teachers. Unbiased estimation of λ would require shrinking teachers' performance toward the school mean rather than toward the grand mean, and using the school mean in place of the grand mean to impute VA predictions to teachers missing leave-two-out VA information. Failure to do so creates downward biases in both $\hat{\lambda}$ and the placebo test coefficients in Table 2, Panel B.

But it is not clear that that this issue is important in practice. The intra-class correlation of teacher VA is 0.2 or less. A correlation of this magnitude is unlikely to cause serious problems if teachers are treated as independent within schools. Appendix B explores alternative VA predictions (e.g., the school mean) for the teachers with missing leave-two-out scores, consistent with different assumptions about the correlation structure. This has essentially no effect on the results.

Finally, it is important to note that excluding teachers with missing VA,

observed in one year but not the other. CFR do not explain the differential treatment.

[26] Other specifications, not reported here, indicate that the significant coefficients in Panel B are – in contrast to the Panel A results – not entirely robust.

as in most of CFR-I's analysis and Panel A of Table 2, relies on auxiliary assumptions as well. The needed assumption here is that there is no sorting of students across classrooms within a school. Since evaluating the extent of such sorting is the entire point of the exercise, one would prefer not to assume it away in estimating λ. Without this assumption, however, the selected-sample estimate $\hat{\lambda}$ is biased toward 1. Moreover, it is clear from Table 2 that ΔQ_{sgmt} is importantly endogenous when computed from the CFR-I subsample. Panel B of Table 2 indicates that the problem is diminished, but perhaps not eliminated, when all classrooms are included.

3.3 Quasi-Experimental Estimates Under A Selection on Observables Assumption

The failure of the placebo test strongly implies that the $\hat{\lambda}$ obtained from the teacher switching analysis, at least as applied to CFR-I's selected sample, is biased upward. The predicted score specification in Table 2, Column 2, suggests that the bias is at least 0.10 in the selected sample, though it may be smaller when all classrooms are included.[27] In Table 3, I explore several approaches to estimating λ without bias.

Panel A follows CFR-I in focusing on the selected subsample of classrooms with non-missing teacher VA predictions. Given the placebo test results, I explore the sensitivity of $\hat{\lambda}$ to the inclusion of controls for the change in student preparedness. Column 1 repeats the specification from Table 1, Column 2. Column 2 adds the change in students' mean prior-year scores as a right-hand side variable.[28] This reduces the $\hat{\lambda}$ coefficient to 0.933 (0.015).

[27] Note that the bias may be larger than the coefficients in Table 2, Column 2 if unobservables change with observables – see footnote 22.

[28] CFR-I present one specification that controls for a cubic in the change in students' mean prior-year scores, in their Table 4, Column 3. This specification also controls for leads and lags of ΔQ_{sgmt}, which are constructed using data from $t-1$ and t so may be endogenous, though coefficients are not reported. In the North Carolina sample, the coefficient on the lead term is highly statistically significant. Taken literally, this is a failed falsification test. But I prefer to exclude the leads and lags of ΔQ_{sgmt}. The result in Column 2 is substantively unchanged when I allow for a nonlinear effect of the mean prior-year score; I focus on the linear model for ease of presentation.

Column 3 presents a specification that excludes the change in prior-year scores but switches the dependent variable to the change in mean residual scores (i.e., to $\Delta\bar{A}_{sgmt}$ rather than $\Delta\bar{A}^{*}_{sgmt}$). This is the specification proposed by CFR-I in developing the quasi-experimental methodology (see their discussion on p. 2617), though in their empirical implementation they use unadjusted scores on the basis of evidence, contradicted above, that changes across cohorts in observable characteristics are orthogonal to ΔQ_{sgmt}. The coefficient here, 0.931, is quite similar to that in Column 2. Column 4 uses the change in gain scores as the dependent variable, as in Figure 3. This yields a somewhat smaller coefficient, 0.889, than in Columns 2 and 3. Note also that each of the methods for controlling for pre-treatment observables yields a more precise estimate than in the unadjusted specification in Column 1 – this added precision is the reason that many experimental analyses control for baseline outcomes even when there is no evidence that the randomization was unsuccessful.

Panel B presents estimates that use all classrooms, assigning teachers observed in only a single year a VA prediction of zero. As noted in Section 3.2, this relies on different, but no less plausible, assumptions than do estimates that exclude such classrooms. Table 1 shows that this simple change, even without controls, reduces the $\hat{\lambda}$ coefficient substantially (from 1.097 to 0.936 in North Carolina data, or from 0.974 to 0.877 in CFR-I's New York sample), and Table 2 showed that the placebo test violation is smaller in this sample. Accordingly, I find that the full-sample $\hat{\lambda}$ coefficient is less sensitive to choices about how to control for student preparedness. Across all four columns, it ranges between 0.83 and 0.90, with standard errors around 0.02.[29]

Appendix Table B1 presents several specifications aimed at testing the robustness of the results to alternative methods of dealing with mechanical relationships between ΔQ_{sgmt} and the change in prior-year scores. Results are quite robust. $\hat{\lambda}$ is near 1 when the selected sample is used without adjustments

[29]The difference between the result in Table 1 and that in Column 1 of Table 3 is that the former reproduces CFR-I's specification, which includes only year fixed effects. Table 3 includes school-year fixed effects in each specification.

for violations of the quasi-experimental design; near 0.93 when the selected sample is used but prior scores are controlled; and 0.86 or a bit smaller when all classrooms are included, with or without controls for additional sorting on observables. These results are not driven by any of the dynamics that CFR (2015a) point to as potential confounding factors. Appendix B presents additional specifications exploring alternative prediction strategies, other than assigning the grand mean, for the teachers excluded from CFR-I's main sample; none have any material impact on the results.

CFR-I present one specification (CFR-I, Table 5, Column 4; reproduced here in Appendix Table A5) that limits the sample to the less than one-third of school-grade-subject-year cells where all of the teachers have non-missing VA predictions, so the issue of sample selection and imputation does not arise. In both New York and North Carolina, the point estimate is roughly similar to the the baseline specification using all cells and including only classrooms with non-missing data. This appears to suggest that sample selection is a non-issue. But these estimates are quite imprecise, given the small sample. More important, CFR-I use a different specification here, including only year effects where their preferred models include school-by-year fixed effects. When school-by-year effects are included in the no-missing-data subsample, results are quite similar to those that I obtain in the full sample. See Appendix Table B3.[30]

I conclude that the best estimate of λ based on the quasi-experimental design, after adjusting for exogeneity failures, is around 0.85. This is near the middle of 0.6-1 range suggested by Rothstein's (2009) simulations, where CFR-I's original results pointed to the very top of that range. Moreover, it indicates a substantively important amount of bias. If we assume that biases are uncorrelated with true effects, $\lambda = 0.85$ implies that $V(b_j)/V(\mu_j) \approx 0.2$. Negative correlations would imply larger bias ratios – a correlation of -0.3 (Horvath, 2015) implies $V(b_j)/V(\mu_j) \approx 0.35$. As I discuss in Section 5, even the smaller estimate is large enough to produce a non-trivial number of

[30]Mansfield (2015) estimates $\hat{\lambda} = 0.832$ when applying the CFR-I strategy to high school teachers' VA and limiting the sample to the no-missing-data subsample.

misclassifications in VA-based evaluations and to create incentives for teachers to manipulate their assignments – by, e.g., refusing to teach classes that will hurt their VA scores – under high-stakes evaluations.

4 Long-Run Effects

The analysis thus far indicates that VA scores are moderately biased by student sorting, with forecast bias around 15% and teacher-level bias of 20-35%. CFR-II's subsequent analysis of the effects of teacher VA on students' longer-run outcomes, such as college graduation or earnings, is predicated on CFR-I's conclusion of unbiasedness. Accordingly, I revisit the CFR-II study here.

CFR-II present two types of analyses of longer-run outcomes. First, for all of the outcomes they consider, they show "cross-class comparisons," simple regressions of class-level mean long-run outcomes on the teacher's predicted VA. Second, for a few outcomes, they also present quasi-experimental analyses akin to those explored above. I reproduce both. I begin in Subsection 4.1 with a discussion of the identification problem and CFR-II's observational strategy. I then present, in Subsection 4.2, estimates of the long-run effects of North Carolina teachers, focusing on the sensitivity to the selection of controls and to the estimation strategy.

4.1 Methods

Following CFR-II, I focus on models for τ_j, the reduced-form impact of a single teacher j on her student's long-run outcomes, not controlling for prior or subsequent teachers. CFR-II's parameter of interest is the covariance between τ_j and the teacher's test score impact, rescaled as $m_j \equiv {}^{\mu_j}/_{\sigma_j}$ where σ_j is the standard deviation of μ_j:

$$\kappa \equiv cov\left(m_j,\, \tau_j\right), \tag{4}$$

Because m_j has unit variance by construction, this is equivalent to the coefficient of a regression of τ_j on m_j. Importantly, while we are interested in the

teacher's causal effect on long-run outcomes, κ is *not* a causal parameter (so does not represent, for example, the effect on long-run outcomes of interventions aimed at raising teachers' test score VA). Rather, it measures the value of VA scores as proxies for teachers' long-run impacts, which even with random assignment would take many years to measure directly.

To estimate κ, CFR-II begin by estimating their VA model using the long-run outcomes in place of end-of-year scores. Paralleling the earlier notation, let Y_i^* represent the outcome for student i, and let $\bar{Y}_{jt}$ be the classroom mean residual after regressing Y_i^* against the VA model covariates, once again using only within-teacher variation. As before, this residual reflects the teacher's true effect τ_j, a bias term b_j^Y that is persistent within teachers, and terms reflecting non-persistent sorting (ν_{jt}^Y) and random variation (e_{jt}^Y):

$$\bar{Y}_{jt} = \tau_j + b_j^Y + \nu_{jt}^Y + e_{jt}^Y.$$

CFR-II estimate κ as the coefficient of a regression of $\bar{Y}_{jt}$ on the standardized predicted test score VA, $\hat{m}_{jt} \equiv {}^{\hat{\mu}_{jt}}/_{\sigma_\mu}$,

$$\hat{\kappa} = \frac{cov\left(\hat{m}_{jt},\, \bar{Y}_{jt}\right)}{V\left(\hat{m}_{jt}\right)} \tag{5}$$

Importantly, though CFR-II refer repeatedly to the inclusion of controls in this analysis, $\hat{\kappa}$ is always estimated via a bivariate regression; covariates are used only to construct the residual long-run outcome $\bar{Y}_{jt}$. This is the reverse of partitioned regression, where the *explanatory* variable is residualized against covariates, and the resulting estimate $\hat{\kappa}$ does not equal the coefficient from a multiple regression of $\bar{Y}_{jt}$ (or Y_i^*) on $\hat{m}_{jt}$ controlling for X_{jt}. CFR (2015a) clarify the reason for this: The parameter of interest here is the coefficient of a bivariate regression of τ_j on μ_j, not the multiple regression coefficient. If students sort to teachers on the basis of τ_j, the covariates X_{jt} might capture some of this sorting, and the multiple regression κ coefficient might understate the value of m_j as a proxy for τ_j.

When the exercise is understood in this way, it is clear that if μ_j and τ_j were

observed directly no exclusion restriction would be required for identification of κ. But neither is observed, and we must rely on the estimates $\hat{\mu}_{jt}$ and $\bar{Y}_{jt}$. This requires assumptions.

First, $\hat{\mu}_{jt}$ must be forecast unbiased, so that the regression of τ_j on $\hat{m}_{jt}$ has the same coefficient as a regression of τ_j on m_j.[31] This is CFR-II's Assumption 1. As discussed above, the evidence suggests that it does not hold.

Second, $\bar{Y}_{jt} - \tau_j = b_j^Y + v_{jt}^Y + e_{jt}^Y$, the estimation error in a teacher's long-run impact, must be orthogonal to the teacher's test score VA $\hat{m}_{jt}$, as otherwise the substitution of the residual outcome $\bar{Y}_{jt}$ in place of the teacher's causal effect τ_j would bias $\hat{\kappa}$.[32] This assumption is problematic as well. Where CFR-I argued that the bias in test score VA (b_j) was likely to be minimal, CFR-II find affirmative evidence that teachers' estimated long-run impacts are biased – that is, that $V\left(b_j^Y\right) > 0$.[33] In this case, the assumption requires that b_j^Y be orthogonal to $\hat{\mu}_{jt}$.

This is untestable, as b_j^Y – reflecting sorting on unmeasured student and family characteristics – is not observed. But the evidence discussed above that measured test score VA is correlated with *observed* family characteristics suggests that it is unlikely to hold. See Appendix Table A7, which shows that teachers with higher predicted VA are assigned students with higher prior scores (included in the VA model) and higher family incomes (not included).

To further illustrate this, Table 4 presents regressions of several student characteristics on the predicted VA of the teacher. Between-school variation is of particular importance, as student socioeconomic status – very strongly predictive of long-run outcomes, but less predictive of annual test score growth – is much more heavily sorted across schools than across classrooms within

[31] We actually require more: The VA forecast error, $m_j - \hat{m}_{jt}$, must be orthogonal to the portion of a teacher's long-run impact that is not captured by her test score VA, $\tau_j - m_j\kappa$.

[32] This is implicit in CFR-II's Assumption 2, which in my notation is that $cov\left(\bar{Y}_{jt} - \kappa\hat{m}_{jt},\, \hat{m}_{jt}\right) = 0$.

[33] See, e.g., CFR-II, p. 2638: "[T]he orthogonality condition required to obtain unbiased forecasts of teachers' earnings VA–that other unobservable determinants of students' earnings are orthogonal to earnings VA estimates–does not hold in practice." See also the the online appendix to CFR-II. In order for long-run VA to be biased but test score VA unbiased, all sorting must be based on unmeasured characteristics that are predictive of long-run outcomes but not predictive of test scores. See the related discussion in Ballou (2012).

schools. Column 1 pools within- and between-school variation; in Column 2, school fixed effects are included so only within-school variation identifies the predicted VA coefficient; and in Column 3, the regressions are estimated on school means to capture between-school variation. Schools with higher average predicted VA teachers have much higher prior year test scores, lower free lunch shares, and higher predicted student outcomes. Within schools, sorting is less dramatic, but teachers with higher predicted VA are statistically significantly less likely to be assigned minority students, students receiving free lunches, and students with lower prior-year scores or predicted end-of-year scores. It thus appears likely that unobserved family characteristics are similarly correlated with $\hat{\mu}_{jt}$, and that the CFR-II strategy confounds the association between τ_j and μ_j with a positive bias term coming from the association of b^Y with $\hat{\mu}_{jt}$.

Below, I show that $\hat{\kappa}$ is is quite sensitive to the inclusion of controls for differences in observed student characteristics across teachers. This strongly suggests that $\hat{\kappa}$ is biased when estimated without controls. But controls for student and family characteristics $\bar{X}_j$ change the estimand from κ to

$$\kappa_X \equiv \frac{cov\left(\mu_j,\, \tau_j \,|\, \bar{X}_j\right)}{V\left(\mu_j \,|\, \bar{X}_j\right)}.$$

This may differ from κ. In particular, if parents and teachers are able to discern teachers' long-run impacts and if they sort on that basis, this would create a causal channel running from τ to $\bar{X}_j$ and imply that $\kappa_X \neq \kappa$.[34] Under this condition, it is exceedingly unlikely for $cov\left(b^Y, \mu_j\right) = 0$, as is required for identification of κ – this would require that the sorting depend only on the part of teachers' long-run effects that is not predictable based on their short run effects, which there is no reason to expect. Thus, even though κ_X may not equal κ, evidence that $\hat{\kappa}_X$ differs from $\hat{\kappa}$ strongly suggests, though does not entirely prove, that $\hat{\kappa}$ is biased relative to κ.

CFR-II also present quasi-experimental analyses of teachers' long-run impacts analogous to those used to estimate forecast bias. I show below that

[34] If students and parents sort to teachers who are known to have high μ_j, but there is no sorting on the basis of $\tau_j - \kappa\mu_j$ (perhaps because it is unknown), then $\kappa_X = \kappa$.

these are as sensitive to the inclusion of controls for observables as are the corresponding short-run quasi-experimental estimates.

4.2 Results

The North Carolina data do not have measures of college enrollment, teen childbearing, or adult earnings, as examined by CFR-II. In their place, I focus on five outcomes that can be measured in 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 are more proximate than CFR-II's outcomes, which mostly measure post-high-school experiences. They also vary in their availability; I focus 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.

Columns 2-4 of Table 5 present observational estimates of κ, from CFR-II in Panel A and from the North Carolina sample in Panel B. The closest alignment between my long-run outcomes and those examined by CFR is for college attendance: I observe self-reported plans as of high school, where CFR-II observe actual enrollment at age 20. The basic observational analysis, in Column 2, indicates that a one standard deviation increase in teacher VA is associated with a 0.82 percentage point increase in the teacher's impact on college enrollment in New York, and with a 0.60 percentage point increase in the teacher's impact on college enrollment plans (and a 1.35 percentage point increase in the impact on four-year college enrollment plans) in North Carolina. I also find positive effects on high school graduation (0.34 percentage point), on high school GPAs (0.022 GPA points), and on class rank (0.54 percentage point). All are highly statistically significant.

Columns 3 and 4 vary the controls used in estimating long-run VA $\bar{Y}_{jt}$, continuing to estimate (5) without controls. In Column 2, the residualization

uses just the covariates from the test score VA model. In Column 3, CFR-II add parental characteristics, drawn from tax returns. These characteristics are not available in the North Carolina data, so I do not repeat these estimates. In New York, their inclusion reduces the estimates of κ by 10-20%, suggesting that bias in $\bar{Y}_{jt}$ that derives from the simpler specification is correlated with $\hat{\mu}_{jt}$. Column 4 replaces the parental characteristics with students' two-years-ago test scores. These estimates are similar to those in Column 3 in New York; in North Carolina, they are mostly smaller than in Column 2, though one (four year college plans) is larger.

Columns 5 and 6 return to the baseline covariates in the construction of $\bar{Y}_{jt}$, but add controls to the second-stage regression of $\bar{Y}_{jt}$ on $\hat{m}_{jt}$. Column 5 uses all of the covariates from the test score VA model, averaged at the teacher-year level; Column 6 further adds teacher-level means of these (aggregating over all of the years that the teacher is observed). All of the $\hat{\kappa}_X$ coefficients are much smaller than the corresponding $\hat{\kappa}$ estimates in Column 2, by 14-45%.[35]

There is every reason to expect that adding the additional family characteristics used in Column 3 (which are not available in the North Carolina data) would lead to additional diminution of the estimated effects. The pattern of results, with sensitivity both to the choice of X_{it} variables in the construction of long-run-outcome VA (Columns 2-4) and to the inclusion of $\bar{X}_{jt}$ variables in the second-stage (Columns 5-6), casts doubt on the interpretation of *any* of the observational estimates as reflecting κ. While this cannot be ruled out – the reduced coefficients in Columns 5-6 of Table 5 could be attributable to differences between κ and κ_X produced by sorting on the sole basis of the portion of teachers' long-run effects that is orthogonal to their test score effects – there is little basis for confidence in the observational model's exclusion restrictions.

Table 6 turns to quasi-experimental estimates of κ. Column 2 reports estimates of the association between the change in mean VA, ΔQ_{sgmt}, and the change in mean unadjusted outcomes, $\Delta \bar{Y}^*_{sgmt}$, as examined by CFR-II.

[35]Responding to an early draft of this comment, CFR (2014c) pointed out that estimates like those in Column 5 and 6 might be biased downward relative to κ_X by measurement error in test score VA. I obtain nearly identical results with a 2SLS estimator that adjusts for measurement error, indicating that this is not an important issue. See Rothstein (2014).

In their preferred specifications, and in contrast to CFR-I, CFR-II include all classrooms in their school-grade-subject-year means, assigning teachers with missing VA predictions the grand mean. I follow that here. Estimates are mostly smaller than the original observational estimates in Table 5, Column 2, and all are much less precise; nevertheless, four of the five are statistically significant. Column 3 adds a control for the change in the mean prior-year score at the school-grade level. Each of the point estimates falls substantially, by at least one-third (and, in the case of the GPA and class rank effects, by over 60%), and none of the adjusted coefficients are significant. When adjusted for observables, the quasi-experimental design offers no evidence that teachers' VA is associated with their long-run effects.

5 Discussion

The first result of my investigation is that essentially all of the empirical results reported by CFR-I and CFR-II from their analysis of New York City students are reproduced, nearly exactly, in data from the North Carolina public schools. Given the dramatic difference in settings, this is remarkable.

But further investigation indicates that CFR's analysis cannot support their conclusions. When I probe CFR-I's test for forecast bias in measured teacher VA, I find that teacher switching does not create a valid quasi-experiment in North Carolina. Measured teacher turnover is associated with changes in student quality, as measured by the students' prior-year scores or just by their demographic characteristics. When changes in observed student quality are controlled, CFR-I's key coefficient $\hat{\lambda}$ is around 0.9, precisely estimated, and highly significantly different from one.

The apparent endogeneity of teacher switching appears to be driven, at least in part, by CFR-I's exclusion of some teachers and classrooms from their quasi-experimental sample. When I include all classrooms, the evidence for endogeneity is weaker, but the forecast bias coefficient falls to around 0.85 and is much less sensitive to the inclusion of controls.

The λ parameter identified by CFR-I's quasi-experiment is only indirectly

related to the quantity of interest, which is the magnitude of biases in individual teachers' VA scores, $V(b_j)$. If one assumes that these biases are orthogonal to teachers' causal effects, my preferred estimate of $\hat{\lambda} = 0.85$ implies that the variance of the portion of student sorting bias that is permanent within teachers (and thus impossible to remove by averaging over several years) is about 18% of the variance of teachers' causal effects. $\hat{\lambda} = 0.9$ would correspond to a variance ratio of 11%. These are roughly in the middle of the range that Rothstein's (2009; 2010) simulations established as consistent with the data.[36] Thus, while CFR-I's strategy narrows the plausible range, it does not support the conclusion that the true value is at one end of that range. Moreover, teacher-level bias is larger if biases are negatively correlated with causal effects (as found by Horvath, 2015; Angrist et al., 2015a). With a correlation of -0.3, teacher-level bias is 24% with $\lambda = 0.9$ and 32% with $\lambda = 0.85$.

To illustrate the potential importance of biases of this magnitude, assume away sampling error – imagine that we observe $\tilde{\mu}_j \equiv \mu_j + b_j$ directly, without error, but that we cannot distinguish the two components. Further suppose that teachers' true effects and the biases in their VA scores are both normally distributed. With $\lambda = 0.85$ and $corr(\mu_j, b_j) = 0$, over one-quarter of teachers with $\tilde{\mu}_j$ in the bottom ten percent will have true causal effects μ_j that are outside the bottom decile.[37] If $corr(\mu_j, b_j) = -0.3$, the misclassification rate rises to over one-third.

This suggests that policies that use VA scores as the basis for personnel decisions will be importantly confounded by differences across teachers in the students that they teach. Teachers with unusual assignments will be rewarded or punished for this under VA-based evaluations. This limits the scope for improving teacher quality through VA-based personnel policies (Rothstein,

[36]CFR-I's VA model is most similar to Rothstein's (2010) "VAM2." A variance ratio of 11% 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 a variance ratio of 18% is quite close to that in Panel C.

[37]In reality, sampling error will also play a role. If decisions are made based on the average of three annual measures of $\tilde{\mu}_j$, each with reliability 0.4 (roughly corresponding to estimates of VA score reliability), nearly half of teachers identified as in the bottom decile will have true μ_js outside of it.

2015). It will also distort teacher assignments as teachers react to the resulting incentive, potentially depressing educational efficiency and offsetting any teacher quality improvements.

Section 4 revisits CFR-II's estimates of the association between teacher VA and teacher effects on students' long-run outcomes. These were in many ways the most important portion of the CFR results, as they suggested that retaining low-VA teachers has extremely important consequences for students' long-run outcomes – that "good teachers create substantial economic value, and VA measures are useful in identifying them" (CFR 2012).

But these results turn out to depend implausible assumptions. CFR-II's "controls" for student observables were implemented in a non-standard way. The conditions required for their estimates to be consistent are quite implausible. Moreover, the estimated long-run effects of high-VA teachers are much smaller when observable differences in students across teachers are controlled directly, both in observational and quasi-experimental analyses. In the more credible quasi-experimental estimates, point estimates are uniformly smaller (more negative) when controls for changes in student observables are controlled, and none are statistically significantly different from zero.

As the North Carolina data have only limited information about family backgrounds and longer-run outcomes, I 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.

Across both investigations, where I am able to estimate the specifications that CFR report, I obtain substantively identical results in the North Carolina sample. CFR have confirmed (in personal communication) that many of my key results obtain in their data, as have Bacher-Hicks et al. (2014) in Los Angeles. It thus seems likely the remainder of my results would generalize across samples as well. The results are also robust to specifications that address a number of objections that CFR (2014e; 2015b) raised in response to an initial draft of this comment, as discussed in the Appendix, which also includes a

rejoinder to CFR's (2015a) Reply.

I conclude that the quasi-experimental methodology proposed by CFR-I, while a major advance in the field, does not support their substantive conclusions. The available evidence suggests that VA scores – in New York, North Carolina, Los Angeles, and likely elsewhere – are moderately biased by student sorting, with a magnitude sufficient to create substantial misclassification rates in VA-based evaluation systems. There is, moreover, no strong basis for conclusions about the long-run effects of high- vs. low-VA teachers, which in the most credible estimates are not distinguishable from zero.

References

Adler, M. (2013): "Findings vs. Interpretation in 'The Long-Term Impacts of Teachers' by Chetty et al." *Education Policy Analysis Archives*, 21.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

American Institutes for Research (2015): "2013-14 Growth Model for Educator Evaluation," Technical report, prepared for the new york state education department, retrieved from `https://www.engageny.org/resource/technical-report-growth-measures-2013-14` on Sept. 25, 2015.

Angrist, J., P. Hull, P. Pathak, and C. Walters (2015a): "Leveraging Lotteries for Value-Added: Bias Reduction vs. Efficiency," Unpublished manuscript.

——— (2015b): "Leveraging Lotteries for Value-Added: Testing and Estimation," Unpublished manuscript.

Bacher-Hicks, A., T. J. Kane, and D. O. Staiger (2014): "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles," Working paper 20657, National Bureau of Economic Research.

Ballou, D. (2012): "Review of "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood"," National Education Policy Center, Boulder, CO,, downloaded Aug. 3, 2015 from `http://nepc.colorado.edu/thinktank/review-long-term-impacts`.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2012): "Great Teaching," *Education Next*, 12.

——— (2014a): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632.

——— (2014b): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–2679.

——— (2014c): "Notes on Imputations and Controls for Observables," Unpublished manuscript.

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

——— (2014e): "Response to Rothstein (2014) on "Revisiting the Impacts of Teachers"," Unpublished manuscript. Downloaded from http://obs.rc.fas.harvard.edu/chetty/Rothstein_response.pdf on October 13, 2014.

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

——— (2015a): "Measuring the Impacts of Teachers: Response to Rothstein (2014)," Unpublished manuscript. Downloaded July 27, 2015 from http://obs.rc.fas.harvard.edu/chetty/va_response.pdf.

——— (2015b): "Measuring the Impacts of Teachers: Response to Rothstein (2014)," Unpublished manuscript. Obtained from AER.

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, 778–820.

Deutsch, J. (2013): "Proposing a Test of the Value-Added Model Using School Lotteries," Unpublished manuscript.

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.

Horvath, H. (2015): "Classroom Assignment Policies and Implications for Teacher Value-Added Estimation," Unpublished manuscript.

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.

Mansfield, R. (2015): "Teacher Quality and Student Inequality," *Journal of Labor Economics*, 33, 751–788.

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

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

——— (2014): "Revisiting the Impacts of Teachers," Unpublished manuscript, October.

——— (2015): "Teacher Quality Policy When Supply Matters," *American Economic Review*, 105, 100–130.

——— (2016): "Revisiting the Impacts of Teachers," Manuscript, March.

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

SAS Institute (2015): "Tennessee Department of Education: Technical Documentation for 2015 TVAAS Analyses," Tech. Rep. Version 1.0, retrieved from `http://tn.gov/assets/entities/education/attachments/tvaas_technical_documentation_2015.pdf` on Sept. 26, 2015.

Value-Added Research Center (undated): "Academic Growth over Time: Technical Report on the LAUSD School-Level AGT Model, Academic Year 2012-2013," Tech. rep., Los Angeles Unified School District, retrieved from `http://achieve.lausd.net/cms/lib08/CA01000043/Centricity/domain/414/documents/AGT%20Informative%20for%202010-2011.pdf` on Sept. 26, 2015.

Revisiting the Impacts of Teachers: Appendix

There are three appendices. Appendix A compares results from each of CFR-I's analyses to those obtained when the analyses are reproduced in the North Carolina data. Appendix B presents alternative specifications aimed at testing for so-called "mechanical" effects and robustness to alternative methods for handling teachers with missing VA predictions. Appendix C responds to CFR's (2015a) critique of my comment.

A Reproduction of CFR-I Results

Appendix Tables A1-A7 present CFR-I's results from New York in parallel with reproductions, using CFR's (2014f) code, in data from North Carolina.

Table A1 presents student-level summary statistics (from CFR-I's Table 1, Panel A). Free lunch and minority shares are lower in North Carolina than in New York, but (surprisingly) the recorded English language learner share is higher. In North Carolina, this variable and special education status are missing from 2009 onward; summary statistics pertain only to those with non-missing data.

Table A2 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 samples 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.

Figure A1 displays the autocorrelations graphically. 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.

Table A3 presents results 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 (0.009) is significantly different from zero where theirs (0.002) is not, but both are small in magnitude. Table A4 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. Again, all estimates are strikingly simi-

lar between the two samples. Table A5 presents estimates from CFR-I's Table 5. Estimates are quite similar, despite the higher share of teachers assigned predicted VA scores of zero in Column 2 in my sample (27.4%) than in CFR-I's (16.4%). Appendix B presents additional relevant results.

Table A6 reproduces CFR-I's Table 6. Notably, the North Carolina results indicate *negative* forecast bias in rows 1-6. But results are generally quite similar.

Finally, Table A7 presents selected estimates from Table 2 in CFR-I's online appendix. 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. In North Carolina, this ratio is 1.36, so coefficients in Panel B are multiplied by this. 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. One notable difference is that minority students have lower-predicted-VA teachers, on average, than non-minority students in North Carolina, but not in New York.

B Additional specifications

B.1 Mechanical effects

Responding to an early draft of this comment, CFR (2014d) suggested that the failure of the placebo test might be due to so-called "mechanical" effects – to factors that influence both prior year scores and measured teacher VA (but perhaps not actual teacher effectiveness). Specifically, CFR note that data from $t-2$ is used both to predict the VA of teachers in $t-1$ and t, and thus to compute ΔQ_{sgmt}, and for the prior-year scores of $t-1$ students. This could create a spurious correlation between ΔQ_{sgmt} and the change in prior year scores. In Table 2 I found that the placebo test failed even when only non-test outcomes were used to measure student preparedness. This demonstrates that test dynamics cannot possibly account for the result. Nevertheless, in Table B1 I explore several alternative specifications aimed at removing the specific mechanical effects that CFR suggest.

Row 1 presents baseline estimates, repeated from Tables 2 and 3. Row 2 is

identical but with standard errors clustered at the school level; this increases standard errors by about one-third.[38]

CFR (2014d; 2015a) suggest that one source of potential mechanical effects is teachers who teach the same cohort 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 both the average VA in grade g in $t-1$ (and thus ΔQ_{sgmt}) and the average lagged scores of grade g students in $t-1$ will reflect her effectiveness.[39] CFR (2014d) propose addressing this by instrumenting for the change in VA, ΔQ_{sgmt}, with a modified measure that excludes teachers who taught $g-1$ in $t-2$ or $t-1$. This is implemented by setting predicted VA for these teachers to zero.

In North Carolina, less than 4% of teacher mobility consists of teachers following students. Not surprisingly, when I modify ΔQ_{sgmt} 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_{sgmt} is correlated 0.96 with the original version, and the first-stage coefficient is 0.98. Estimates of my key specifications are shown in Row 3 of Table B1. When classrooms with missing VA scores are excluded, the association with the change in prior-year scores is reduced but remains significant, and the λ estimate is hardly changed. Note that the no-follower instrument involves setting some teachers' VA predictions to the grand mean, and thus relies on the same assumption of within-school independence as does the inclusion of teachers with missing leave-two-out predictions, also set to the grand mean. There is thus no set of assumptions that can justify the subsample specifications in columns 1-3. When all classrooms are included, in columns 4-6, the placebo test coefficient is no longer significant, but the λ coefficient from a specification without controls falls to match that in the specification with controls. I thus conclude that "follower" teachers might contribute slightly to the placebo test violation, but that recognition of this phenomenon has no effect

[38]CFR-I's main results cluster at the school-by-cohort level. School-level clustering is more general. Moreover, I present below IV specifications with school-year fixed effects; it is computationally difficult to cluster these at the school-cohort level.

[39]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. Note also that "following" is a problem for the quasi-experimental analysis as well as for the placebo test. The quasi-experimental 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 $\hat{\lambda}$ 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.

on my conclusions regarding forecast bias.[40]

CFR (2014d; 2015a) also suggest that school-year-subject shocks could create mechanical, spurious failures of the placebo test: A positive shock to a school in $t-2$ will raise both the predicted VA of the school's $t-1$ teachers and the prior-year scores of the $t-1$ students. This would be absorbed by school-year effects already included in the main specifications if it were common across subjects, but subject-specific shocks would not be. CFR (2014d; 2015a) propose to address it by including school-subject-year fixed effects. I implement this in Row 4. This halves the number of degrees of freedom, leaving only three or fewer observations per cell. Standard errors are larger here. The quasi-experimental estimates in Columns 2 and 3 rise, and I cannot reject $\lambda = 1$ in Column 3. However, in the preferred sample that includes all classrooms (assigning VA predictions of zero to teachers with missing data), the additional fixed effects make little difference at all, and I decisively reject $\lambda = 1$. Row 5 presents a specification with both school-subject-year effects and instrumentation for follower teachers. The main placebo test coefficient is insignificant here, but my preferred forecast bias coefficient (in column 6) is unchanged, at 0.89, and remains significantly different from 1.

The inclusion of school-subject-year effects is not the only way to address the possibility that common shocks would affect both teachers' VA predictions and students' lagged scores. An alternative, more consistent with the overall research design, is to exclude $t-2$ data from the predictions of teacher VA in years $t-1$ and t. "Leave-three-out" VA predictions, ensure that there is zero overlap between the scores used to construct the VA scores and those used for the dependent variable in the placebo test, as the latter is based only on data from $t-2$ and $t-1$. Row 6 presents estimates using these leave-three-out VA predictions. They are quite similar to the baseline estimates, if anything indicating larger selection problems and smaller quasi-experimental estimates. Row 7 combines the leave-three-out VA scores with the no-follower IV, with quite similar results

CFR (2015a) point out that with serial correlation in the school-year-

[40]I have also explored specifications analogous to those in Columns 3 and 6 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. When CFR (2015a) estimate the specification in Column 1, the coefficient is insignificantly different from zero, though this coefficient is significant in Los Angeles (Bacher-Hicks et al., 2014). This may be the sole substantively important difference in empirical results across the three samples. In any event, when CFR (2015a) use the "no followers" design for the main quasi-experimental specification (as in Column 2), they estimate $\hat{\lambda} = 0.92$ and reject the null hypothesis that $\lambda = 1$. This is quite similar to my results.

subject shocks, a shock in $t-3$ would influence leave-three-out VA scores and be correlated with the shock to prior-year scores for the $t-1$ cohort, potentially biasing leave-threee-out placebo test. Such serial correlation would create a similar bias in the CFR-I quasi-experiment, as $t-2$ shocks enter into VA scores and would be similarly correlated with the shock to $t-1$ scores, and indeed one would expect the leave-three-out strategy to reduce bias.

Nevertheless, rows 8 and 9 present estimates that use leave-four-out and leave-five-out VA scores that exclude not just $t-2$ but also $t-3$ and (in Row 9) $t-4$ data from the calculations. Results are extremely stable. In row 10, I take this to the logical extreme, using only data from $t+1$ and thereafter to forecast (backcast) VA in $t-1$ and t. This specification, proposed by CFR (2014d), should entirely eliminate any mechanical effect of the form that CFR (2014d; 2015a) propose, but estimates are basically unchanged – if anything, the forecast bias coefficient falls from the baseline specification ($\hat{\lambda} = 0.83$ vs. 0.86).

Taking the various specifications in Table B1 together, along with the non-test placebo analysis in Table 2, the evidence is clear that mechanical effects cannot account for the results.

B.2 Teachers with missing leave-two-out predictions

CFR-I's key VA measure used in each paper is a "leave-two-out" forecast of a teacher's outcomes in year t or $t-1$ based only on data from prior to $t-1$ or after t. This forecast can be seen as an Empirical Bayes prediction of the teacher's impact in $t-1$ or t, and by construction is an unbiased prediction of the VA score in that year. When teachers are observed only in $t-1$ or t, however, there is no other data on which to base this forecast. In most of their analyses, CFR-I exclude such teachers, and their students, from their calculation of school-grade-year means. I argue above that this sample selection biases the key coefficient $\hat{\lambda}$ toward the null hypothesis of $\lambda = 1$. Following one specification in CFR-I and most of the analysis in CFR (2014b; "CFR-II"), he includes these teachers and their classrooms, assigning them a VA prediction equal to the grand mean.

The grand mean is an unbiased prediction of every teacher's VA, and is the logical extension of the Empirical Bayes methodology for CFR-I's leave-two-out predictions. But the relevant prediction for CFR-I's quasi-experimental analysis is of the school-grade-year mean VA, not that of the individual teacher. If VA is correlated across teachers within schools, then the average of unbiased forecasts for each teacher is a biased forecast of the average VA at the

school. Failure to account for this would create upward bias in both CFR-I's quasi-experimental coefficient $\hat{\lambda}$ and my placebo test coefficient. Importantly, this bias arises even if leave-two-out forecasts are available for every teacher. Avoiding it would require shrinking teachers' observed performance toward the school mean rather than toward the grand mean, and using school average performance rather than the overall average to predict VA for teachers with missing leave-two-out data.

Table B2 explores alternative strategies for assigning VA predictions to teachers with missing leave-two-out data. Following CFR (2015a), I use CFR-I's leave-two-out predictions for teachers for whom they are available in every specification in this table, though the above discussion suggests that the should be changed as well.

Panel A presents CFR-I's main regression of the year-over-year change in school-grade-subject mean test scores on the corresponding change in mean teacher predicted VA. Panel B presents my placebo test, replacing the dependent variable with the change in mean *prior year* scores. Panel C augments the Panel A specification with a control for the change in mean prior year scores.

The first two columns reproduce estimates from the main paper for context: Column 1 leaves the teachers with missing leave-two-out predictions and their students out of the school-grade-year means, while column 2 includes them using the grand mean for the teachers' VA predictions. When the teachers are left out, $\hat{\lambda} = 1.03$ (standard error 0.02) when students' prior scores are not controlled, and the null hypothesis of $\lambda = 1$ is not rejected. But the placebo test fails, with a highly significant coefficient of 0.14, and when students' prior-year scores are controlled the key coefficient falls to 0.93 (0.02) and the null hypothesis is rejected. When teachers with missing leave-two-out predictions are included, even the baseline specification in Panel A rejects the null hypothesis ($\hat{\lambda} = 0.90$, SE 0.02). The placebo test result is weaker but still significant, and the specification that controls for observables yields $\hat{\lambda} = 0.86$ (SE 0.02).

Columns 3-5 present results from other imputations. Column 3 uses the (appropriately shrunken) mean residual of all teachers at the school in all years other than $t-1$ or t to forecast the VA of teachers in those years who are not seen outside that window. This method would be robust to correlations among teachers at the same school. Column 4 uses the mean residual of all teachers across all schools who are observed for two years or less. This captures the possibility that the teachers with missing leave-two-out predictions may systematically differ from others. Finally, Column 5 uses the mean for such teachers at the same school, as in other cases using only data from outside the

$t-1$ to t window.

Results are qualitatively similar across all of the different imputation models. In each case, the baseline specification in Panel A yields an estimated $\hat{\lambda}$ between 0.90 and 0.93, all significantly different from one. The placebo test fails regardness of the imputation used, with the models that use only same-school data indicating much larger placebo test violations. And when prior scores are controlled, the key coefficient falls to between 0.85 and 0.89, again always significantly different from one. It is clear that non-independence of teacher VA within schools cannot account for my results.

Table B3 takes a different approach to the issue of missing leave-two-out predictions. Column 2 of CFR-I's Table 5 suggests a substantial degree of forecast bias when teachers with missing VA predictions are assigned the grand mean VA, and as Table 1 indicates the same is true in the North Carolina sample. But CFR (2015a) point instead to Columns 3 and 4 of CFR-I, Table 5, reproduced for the North Carolina sample in Table A5. These limit the sample to school-grade-subject-year cells with few (Column 3) or no (Column 4) missing VA predictions, and in each sample they indicate less forecast bias. CFR (2015a) interpret this as evidence that the imputation algorithm accounts for the result in Column 2, and argue that the Column 4 result in particular indicates that VA predictions are unbiased, at least in the subsample of school-grade-subject-year cells with no missing VA predictions.

But this result is not at all robust. In particular, it evaporates when school-year fixed effects are added. These fixed effects are included in CFR-I's main specifications but omitted without explanation from their Table 5.

The odd numbered columns of Table B3 report the four specifications from CFR-I's Table 5. Note that the placebo test coefficients are quite large in these columns, though the models with controls in columns 1, 5, and 7 yield λ estimates that are not distinguishable from 1 (in large part because the models without controls yield λ estimates well in excess of 1).

As noted, these specifications, following CFR-I, include only year fixed effects, rather than the school-year effects included in the models that CFR-I prefer in the rest of their analysis. This raises the possibility of bias from unmodeled school trends. The even numbered columns of Table B3 add back the school-year fixed effects.[41] This change reduces the placebo coefficients, which become insignificant in columns 6 and 8. But it also reduces the forecast bias coefficients. CFR-I's preferred model, which limits the sample to cells with

[41] One might worry that the no-missing-data subsample in Column 7 is not large enough to permit any degree of precision with school-year fixed effects. But standard errors increase by less than 20% when these are added, much less than the increase (of nearly 100%) when cells with missing VA predictions are discarded.

no missing data, yields a forecast bias coefficient of $\hat{\lambda} = 0.92$ without controls and 0.90 (significantly different from one) with a control for the change in prior year scores. This is broadly similar to what is obtained from the full sample.

C Rejoinder to CFR (2015a)

The exchange between myself and Chetty, Friedman, and Rockoff (CFR) has involved several rounds of private communication, dating back to 2010, and a more recent exchange of public drafts and responses. Throughout, it has been constructive and scholarly, and I have learned a great deal from it. I am grateful to CFR for their role in it, and the current draft of my Comment (dated March 2016) reflects many good points that CFR have made.

Nevertheless, CFR and I continue to have sharply different interpretations of what the empirical patterns mean for the substantive questions under investigation. My Comment reflects my interpretation; CFR offer a very different interpretation in their Reply. In this appendix, I discuss the July 2015 version of CFR's Reply (CFR 2015a), written in response to the October 2014 version of my Comment (Rothstein, 2014). CFR may update their Reply to respond to the revised version of my Comment. If so, I will update this rejoinder. To ensure a complete record, the original rejoinder (dated March 2016) will remain posted on my webpage, at `http://eml.berkeley.edu/~jrothst/CFR/supplement_mar2016.pdf`.

I respectfully disagree with many of the conclusions drawn by CFR (2015a), which in many cases are based on claims that are theoretically correct but turn out, upon investigation, to be empirically irrelevant. None of the evidence presented by CFR (2015a) alters the main conclusions of my earlier draft, which persist in the current version:

1. That the CFR-I (2014a) research design is not a valid quasi-experiment because the treatment is correlated with observable determinants of the outcome;

2. That much but not all of the problem derives from CFR-I's exclusion of a non-random subset of classrooms from school-grade-subject-year means;

3. That estimates that adjust for differences in observables indicate a non-trivial but not enormous degree of "forecast bias"; and

4. That estimates of teachers' long-run effects are not at all robust and quite likely to be biased by student sorting.

I begin by laying out CFR (2015a)'s six main arguments, in order of their importance to my conclusions, along with my responses. I follow this by presenting simulation evidence to support one of these responses. In the interests of space, I do not discuss other arguments made in CFR's response that are less relevant to my conclusions.

CFR (2015a)'s six main arguments are:

1. *Examination of prior test scores is not informative about the validity of CFR-I's quasi-experimental research design, because value-added is estimated from prior test scores and is thus mechanically correlated with them.*

It is theoretically correct that the use of prior test scores in the construction of the VA measures could create a spurious correlation, making it appear that changes in teacher VA are not randomly assigned. But in practice, this does not account for the result. The main text and Appendix B present a number of analyses that probe this possibility. All indicate that the failure of the placebo test is real, not spurious. The most definitive is an alternative placebo test that is based solely on non-test student characteristics (race, gender, special education, free lunch status, limited English status, grade repetition, etc.). This test is entirely immune from mechanical correlations, but also shows that changes in mean teacher VA, as estimated by CFR-I, are significantly related to changes in student preparedness (see Table 2[42]).

2. *The primary source of the correlation between changes in teacher value added (VA) and changes in prior test scores is common shocks that affect both. When these so-called "mechanical effects" are addressed via changes in the specification, the correlation is eliminated.*

CFR (2014d; 2014e; 2015a) have advanced this idea in a series of public responses over the last eighteen months, pointing to potential mechanical effects

[42]Unless otherwise specified, all table references are to tables in the March 2016 version of my comment, Rothstein (2016).

deriving from teachers who follow students across grades or from school-year-subject-level shocks. As noted above, explanations based on test score dynamics cannot possibly account for the placebo test result, as it holds even when non-test variables are used in place of prior test scores. Moreover, for each proposed mechanical channel, I have implemented alternative specifications of the placebo test that close off that channel. In particular, I close off the teacher-follower channel by instrumenting with VA changes computed only over non-follower teachers, and I close off the school-year-subject shock channel by using "leave three out" VA measures that do not rely on data from $t-2$ in computing VA predictions for $t-1$ or t. Results are remarkably stable across specifications (see Appendix Table B1).

CFR (2015a) suggest that there may be school-level shocks that are correlated across years, so that shocks in $t-3$ influence both VA predictions for $t-1$ teachers (even when $t-2$ data are excluded) and the prior year scores of $t-1$ students, which are measured in $t-2$. Serially correlated school-level shocks could produce the failure of my placebo test even when I use leave-three-out VA scores that do not rely on $t-2$ data.

To ensure that my results are not driven by this channel, I estimated specifications that exclude all data from several years before the $\{t-1, t\}$ window from the VA predictions. If in fact the placebo test result derived from serially correlated shocks, the coefficient should decline as more years are excluded. But in fact this has essentially no effect on the results – even when I base VA predictions solely on *future* data. Thus, while CFR-I present simulation evidence that serially correlated shocks *could* drive the results, the empirical evidence from real data indicates that they do not.

It is also worth noting that the dynamics that CFR (2015a) propose as sources of mechanical effects would in general invalidate not just the placebo test but also CFR-I's quasi-experimental research design itself, and would lead CFR-I to understate forecast bias. School-year or school-subject-year shocks that are correlated between $t-2$ and $t-1$ would invalidate the design, as the leave-two-out teacher VA predictions for $t-1$ would be influenced by shocks correlated with those to students' $t-1$ test scores.[43] It would take a very particular dynamic structure to generate correlations between $t-3$ and $t-2$ scores but not between those in $t-2$ and $t-1$. Similarly, the presence

[43]CFR (2015a) present a specification with school-subject-year FEs. But with only two or three observations (grades) per school-subject-year cell, these specifications rely very heavily on a strict exogeneity assumption that is prima facie violated by teachers who switch grades within schools. In my explorations with simulated data – including with the data generating process of the simulations used in CFR (2015a)'s Table 4 – I have found that these specifications are very poorly behaved.

of meaningful numbers of "follower" teachers would imply that the outcome in the quasi-experiment reflects not only the quality of the grade-g teachers but also the (correlated) quality of grade $g-1$ teachers, and thus that the quasi-experimental coefficient overstates the parameter of interest, λ.

3. *The augmented quasi-experimental specification that includes a control for the change in prior year scores yields a biased estimate of the forecast bias coefficient λ.*

Again, this is theoretically possible, but the claim that it is relevant in practice is pure speculation unsupported by evidence. CFR (2015a) hypothesize that the change in prior year scores has two components, with one component correlated with the change in VA but not with the change in end-of-year scores and the other correlated with end-of-year scores but not with VA. This might be a reasonable hypothesis if the "mechanical effects" claims discussed above held up. Even here, quite restrictive dynamic structures would be needed to generate mechanical effects from sources that are uncorrelated with the dependent variable in CFR-I's analyses. CFR (2015a) argue for "nonparametric" specifications, but their specifications and simulations generally rely on quite strong implicit assumptions. But as noted above, the evidence does not support CFR's claims about mechanical effects. Without them, while anything is possible, the only reasonable conclusion is that CFR's (2015a) conclusions rely on quite speculative, unsupported assumptions.

It is also possible, and more likely, that both the specification without a control for prior year scores (as in CFR-I) and one with such a control (as in my preferred analyses) are biased by unmeasured components of the endogeneity of teacher VA changes. I do not claim that the specification with controls is highly credible. But in the presence of clear evidence that the quasi-experimental treatment is not randomly assigned, and that this is *not* attributable to CFR (2015a)'s hypothesized mechanical effects, a specification with controls is preferable, in my view, to one that does nothing to address the endogeneity of treatment. Moreover, I show (see Table 3) that the top-line result of forecast bias around 10-15% (i.e., of $\hat{\lambda}$ around 0.85-0.9) is robust to several ways of addressing the endogeneity, which adds to my confidence in the result.

4. *An analysis restricted to school-grade-subject-year cells without missing data is the most definitive way to address concerns about sample selection*

due to missing data, and validates CFR-I's conclusion that VA scores are forecast unbiased.

I disagree that this is the most definitive way to address concerns about sample selection due to missing data – it requires discarding between three-quarters (New York) and four-fifths (North Carolina) of the school-grade-subject-year cells, and estimates are quite imprecise. Moreover, the remaining sample includes fewer teachers who are new to teaching or to the sample grades, and forecast bias in this subsample might be different from that in the broader population.

More importantly, as discussed in Section B.2, above, the subsample analysis does not validate the conclusion of no forecast bias. First, I find that the placebo test coefficient is quite large and statistically significant even in the complete data subsample. Second, CFR-I inexplicably drop the school-year fixed effects from their preferred specification when they analyze the complete data subsample. When I include them the estimate of λ is 0.918 without controlling for prior year scores and 0.899 (and significantly different from one) when this control is included. This is broadly similar to what is obtained from the full sample.

Thus, at most this subsample analysis shows that not *all* of the problem with CFR-I's specification is attributable to their exclusion of a non-random subset of classrooms from school-grade-subject-year means. It does not demonstrate (or even point in the direction) that the design is valid, or that forecast bias is zero, even locally for the small subset of schools without missing data. CFR (2015a)'s statement that "[t]his approach consistently yields estimates of forecast bias close to zero in both the CFR and North Carolina datasets" is incorrect as it applies to North Carolina, and the single specification that CFR have reported from their dataset is not enough to demonstrate the point there either.

5. *The inclusion of all classrooms in the analysis, using grand mean imputation, generates downward-biased estimates of the key parameter λ.*

We are in agreement that analyses that include all classrooms are not definitive, but rest on the appropriateness of the model used to predict teachers' VA. I focus on specifications that use the grand mean because this is the strategy proposed by CFR, who use it throughout their analyses for some (most of CFR-I's specifications) or all (one failed robustness test in CFR-I, and the

main specifications of CFR-II) of the classrooms with missing data.[44] It is also consistent with CFR's prediction model (seen as an example of Empirical Bayes methods) for classrooms that have data.

That said, the claim that my use of grand mean predictions accounts for my results is incorrect. CFR (2015a) are correct that positively correlated VA across teachers within schools could lead to attenuation with grand mean predictions.[45] But again, this theoretical point is not empirically relevant. Results of both the placebo test and the forecast bias estimation are robust to a variety of alternative prediction strategies, including some that are robust to non-independence of teacher VA within schools (which is the source of bias under grand mean predictions). See the discussion in Section B.2, above. And even when I follow CFR-I's preferred strategy of excluding classrooms without teacher VA predictions, the results are quite clear that λ is less than one in any specification that does anything to address the endogeneity of changes in teacher VA (Table 3).

Four other points are worth noting about the imputation issue:

- CFR (2015a)'s attenuation argument may help to explain why some of the placebo test coefficients are smaller when all classrooms are included than when they are not (see Table 2); it suggests that the failure to reject the placebo test null hypothesis in some all-classroom specifications should not be taken as support for the exclusion restriction.

- CFR (2015a) present a simulation to demonstrate the bias from the grand mean imputation, but this uses a counterfactually large intra-school correlation of teacher VA ($\rho = 0.35$). When I use a value that is empirically grounded ($\rho = 0.2$), the bias in the simulations is quite small. CFR's (2015a) simulation is explored below in subsection C.1.

[44] Throughout all of their quasi-experimental analyses, CFR-I and CFR-II impute VA scores of zero for teachers observed in $t-1$ and t but not in other years. At issue is whether to apply the same imputation to teachers observed only in a single year, as is done in CFR-I's Table 5, Column 2 and throughout CFR-II, or to exclude these teachers and their students from the analysis, as is done elsewhere in CFR-I. I see no basis for viewing the grand mean as the correct prediction for the first group of teachers but not for the second, and CFR have never offered an explanation for this.

[45] They are also correct that using all classrooms on one side of the regression and a subset on the other can lead to biases. An earlier draft of my comment (Rothstein, 2014) presented estimates of this form to build intuition for the full-sample results. CFR (2015a) quite reasonably objected that these specifications were not very informative. They have therefore been removed.

- CFR's simulation assumes that there are no differences across classrooms in students' prior achievement. My argument for the importance of accounting for classrooms with missing teacher VA was predicated on the empirical result that students' prior scores are positively correlated with teacher VA, so excluding a classroom has effects of the same sign on mean teacher VA and mean student preparedness that bias the $\hat{\lambda}$ coefficient upward. It is thus not surprising that CFR's simulation shows no bias from excluding classrooms with missing VA, as it fails to include the relevant features of the real data. Where the real data are concerned, CFR (2015a) may object to the particular imputation model proposed by CFR-I, but they do not dispute that excluding classrooms with missing data, as in CFR-I's main analyses, biases $\hat{\lambda}$.

- Finally, the data generating process for CFR (2015a)'s simulation violates the exclusion restrictions that CFR-I require to identify λ, even with random assignment and complete data, as these restrictions rule out non-zero intra-school correlations. If the intra-school correlation is non-zero, the change in the average of unbiased predictions of individual teachers' VA is not an unbiased prediction of the change in the average VA. If the correlation is positive, CFR-I's methods will likely overstate the change in VA, biasing $\hat{\lambda}$ upward. This could offset bias from endogenous teacher switching (or from endogenous sample selection).

These points are discussed in more detail in Section C.1, below.

One final point: While we agree that specifications that include all classrooms rest on the appropriateness of the model used to predict teachers' VA, it is also true that specifications, like those that CFR-I prefer, which exclude a non-random set of classrooms also rest on assumptions. These assumptions are quite implausible – they require that student preparedness be uncorrelated with teacher VA. It is empirically the case that students' observables *are* correlated with teacher VA; whether their unobservables are as well is the entire point of the CFR-I exercise. So while it is reasonable to disbelieve specifications that rely on imputations, it is not reasonable to treat those that simply exclude teachers with missing data as unbiased.

6. *It is not the case that a regression of long-run outcomes on teachers' test score VA, with controls for observables, is consistent under more general conditions than is CFR-II's two-step procedure.*

This point responds to an earlier version of my comment (Rothstein, 2014). CFR (2015a)'s discussion of this issue clarified it substantially for me, and the revised comment has been rewritten with this in mind.[46] I believe that the main point stands.

CFR are correct that the exclusion restrictions under which my approach identifies κ do not strictly nest those under which CFR-II's approach identifies that parameter, and that when students sort into classrooms on the basis of teachers' impacts on long-run outcomes (i.e., on the basis of τ_j) then their approach can be consistent for κ even when mine is not. Nevertheless, I remain unconvinced that their exclusion restrictions are remotely plausible.

A useful way to see it is that regressions with controls identify a potentially different parameter, κ_X, under weaker – still not very plausible, but more so – restrictions. The two parameters are equal unless students are sorted into classrooms on the basis of the portion of teachers' long-run effects that cannot be predicted by the teachers' test score value added. I view this kind of sorting as implausible – I think it unlikely that parents can discern teachers' long-run impacts – so I think the parameters are likely to be quite similar, and I view the difference between the $\hat{\kappa}$ and $\hat{\kappa}_X$ estimates as a sign that the former is biased due to failures of CFR-II's exclusion restrictions.

One may or may not interpret $\hat{\kappa}_X$ as a good estimate of κ_X. But the evidence clearly indicates that the conditions required for CFR-II's approach are not satisfied. Thus, we do not have reliable estimates of κ. In my view, the fact that results are quite different under my approach is a strong indication, though not definitive proof, that the CFR-II strategy overstates teachers' long run impacts by a great deal.

C.1 Simulations of the effect of missing data

Under point 5, above, I referred to CFR's (2015a) simulation evidence about the effect of different ways of handling teachers with missing VA predictions. In CFR's simulation, VA is unbiased – indeed, it is measured without any error at all. Thus, the true value of λ is one. CFR (2015a) show that in this case, $\hat{\lambda}$ is close to one when data are available for all teachers or when

[46] In personal communication regarding the long-run analysis, CFR emphasized measurement error in teacher VA. Responding to this, I (Rothstein, 2014) presented IV specifications designed to eliminate attenuation due to measurement error in an explanatory variable, with zero impact on the results. CFR now point to a different dynamic, so I no longer emphasize the IV results.

teachers with missing data are excluded from the analysis, but that $\hat{\lambda}$ is only 0.88 when teachers with missing data are included with their predicted VA scores set to zero. This last result is driven by an assumption that VA is positively correlated among teachers in the same school; failing to account for this in assigning VA predictions to teachers without them leads to overstating the magnitude of changes in VA.

But there are two big problems with this simulation. First, the intra-class correlation (ICC) in the simulation is set to 0.35, which is far too large. CFR (2015a) report that the ICC in the actual New York data is only 0.2; I obtain a somewhat smaller value, around 0.16, in North Carolina. An ICC of this magnitude does not cause much of a problem for the grand mean predictions. Table B4 reproduces CFR (2015a)'s simulation results in row 1, then reports results using a more realistic ICC of 0.2 in row 2. With grand mean predictions, $\hat{\lambda} = 0.93$, much closer to one than in the large-ICC simulation or than in the empirical results from either the New York or the North Carolina samples.

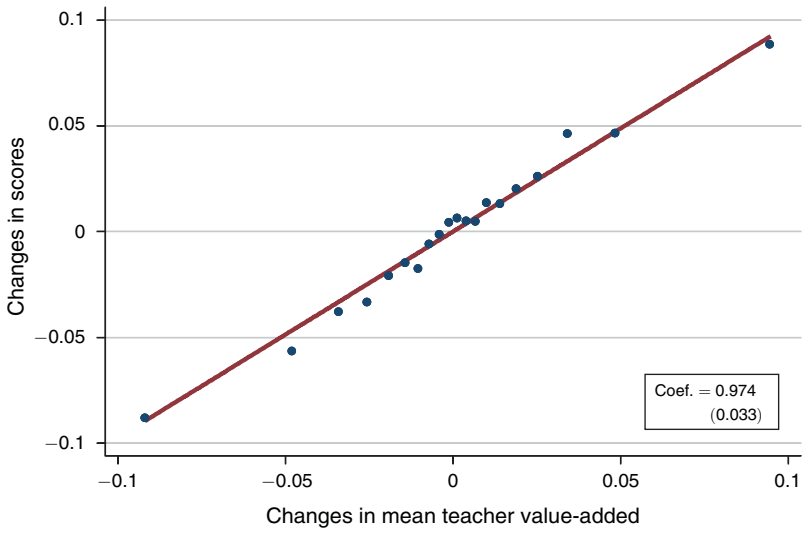
Second, CFR (2015a)'s simulation assumes that teachers' VA is known with certainty. In fact, a key portion of the CFR-I empirical strategy is to predict each teacher's VA in one year based on noisy measures of her performance in other years, and these predictions assume the ICC is zero. With a non-zero ICC, CFR-I's methods do not identify the degree of forecast bias.[47] Rows 3 and 4 of Table B4 extend the CFR (2015a) simulation to include predictions of VA scores based on observed outcomes in other years. I assume that each teacher is observed in four years other than the ones used for the quasi-experimental analysis, and that each year provides an independent noisy signal of the teacher's underlying VA with reliability 0.4. I do not allow drift in teacher quality across years. I use a high ICC of 0.35 in Row 3, and a lower value of 0.2 in Row 4. These simulations yield estimates of λ that are well below one (0.86 and 0.93, respectively) even when VA predictions are available for all teachers. This suggests that with a positive ICC, an estimate of $\hat{\lambda} = 1$ will obtain only if $\hat{\lambda}$ is upward biased from some other source, such as an association between ΔQ_{sgmt} and the change in prior determinants of student outcomes.

[47] Specifically, CFR-I construct ΔQ_{sgmt} as the change in the average of unbiased predictions (if $\lambda = 1$) of teachers' VA scores. But their Assumption 3 requires that ΔQ_{sgmt} be an unbiased predictor of the change in the average true VA. When the ICC is not zero, the average of unbiased predictions is not an unbiased prediction of the average. Thus, a non-zero ICC implies that CFR-I's Assumption 3 is violated, and thus the b coefficient from CFR-I's equation (15) does not identify λ. CFR (2015a)'s characterization of their simulation ("simulated data in which none of CFR's identification assumptions are violated") is therefore incorrect.

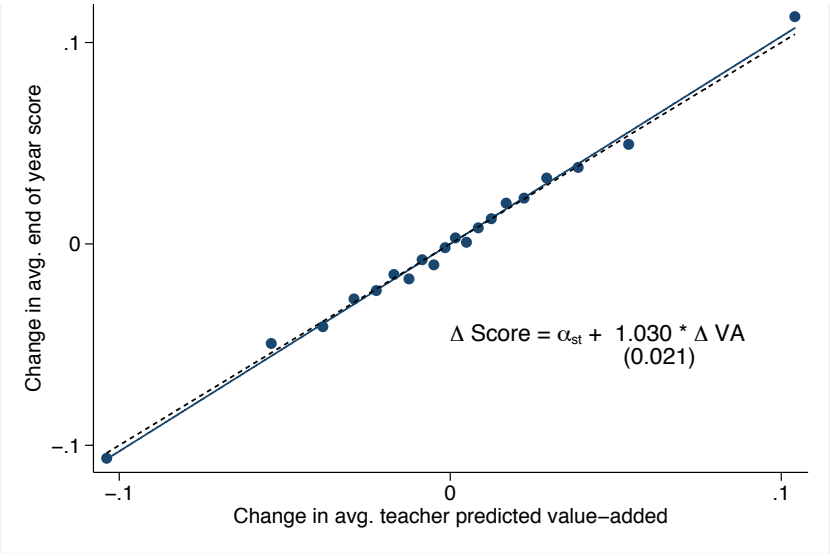
In other words, it is odd that CFR (2015a) defend their methods by pointing to the inappropriateness of grand mean imputation in the presence of a correlation among teachers at the same school, as (a) CFR-I use exactly this imputation for many teachers throughout their analysis and (b) CFR-I's entire empirical strategy is predicated on an (implicit) assumption that this correlation is zero. Moreover, in CFR (2015a)'s own simulation an empirically reasonable value of the ICC does not lead to enough attenuation to account for the empirical results.

Figure 1
Bin-Scatter Plot of Change in Average Teacher Predicted VA and Change in Average End-of-Year Score

Panel A: New York City

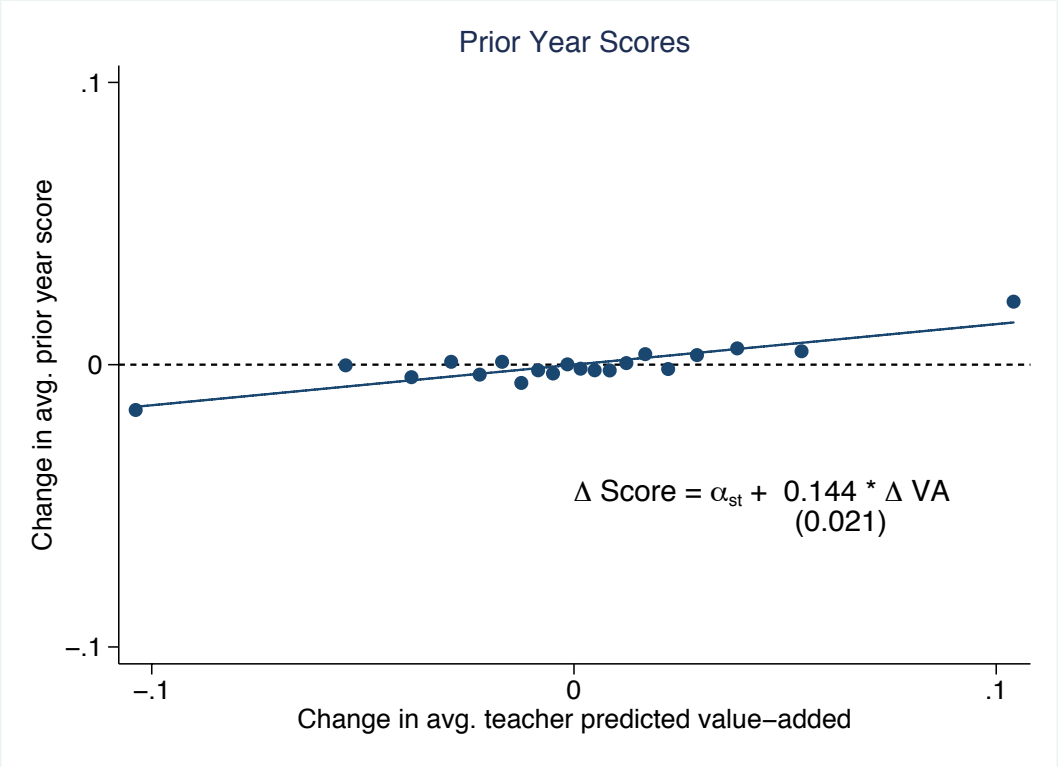


Panel B: North Carolina



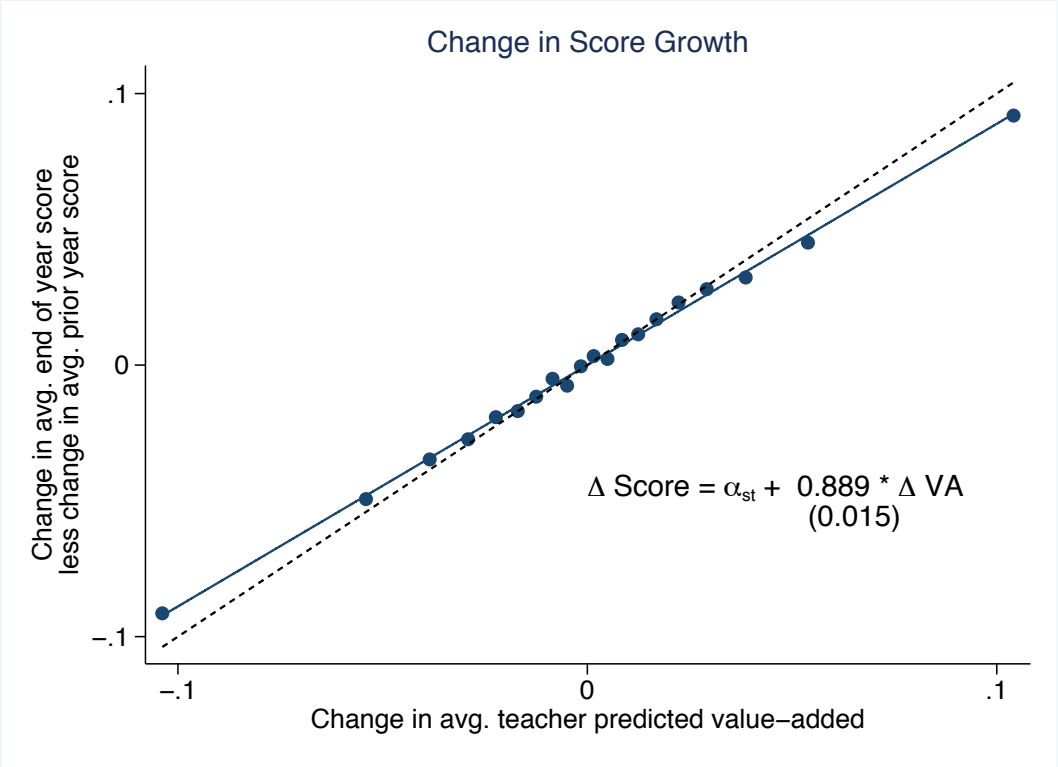
Notes: Panel A is taken from CFR-I, Figure 4A, and corresponds to Table 1, Column 1, Panel A. Panel B is constructed similarly using North Carolina data and corresponds to the sample used in Table 1, Column 2, Panel B. Each presents a binned scatter plot of cohort-to-cohort changes in school-grade-year-subject average scores against changes in school-grade-year-subject average predicted teacher VA, after residualizing each against year (Panel A) or school-year (Panel B) fixed effects. 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 year (panel A) or school-year (panel B) fixed effects; coefficients and standard errors (clustered at the school-cohort level) are shown on each plot.

Figure 2
Bin-Scatter Plot of Change in Average Teacher Predicted VA and Change in Average Prior Year Score



Notes: Figure is identical to Figure 1, Panel B, except that the variable plotted on the vertical axis is the mean cohort-over-cohort change in prior-year (rather than end-of-year) scores in the vingtile group. Sample and regression equation correspond to Table 2, Column 1, Panel A.

Figure 3
Bin-Scatter Plot of Change in Average Teacher Predicted VA and Change in Average Gain Score



Notes: Figure is identical to Figure 1, Panel B, except that the variable plotted on the vertical axis is the mean cohort-over-cohort change in gain scores (the student-level growth in scores from the end of one year to the end of the next) in the vingtile group. Sample and regression equation correspond to Table 3, Column 4, Panel A.

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

Dependent variable:	Δ Score	Δ Score	Δ Score (Predicted)	Δ Score (all students)
	(1)	(2)	(3)	(4)
	Panel A: CFR (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
	Panel B: North Carolina reproduction			
Change in mean teacher predicted VA across cohorts	1.097	1.030	0.008	
	(0.022)	(0.021)	(0.011)	
Change in mean teacher predicted VA across cohorts (with zeros)				0.936
				(0.022)
Year fixed effects	X			X
School x year fixed effects		X	X	
Grades	3 to 5	3 to 5	3 to 5	3 to 5
# of school x grade x subject x year cells	79,466	79,466	54,663	91,221

Notes: Panel A is taken from the indicated Tables and Columns of CFR (2014a); Panel B is estimated using the same variable construction and specifications in the North Carolina sample. The dependent variable in each column is the year-over-year change in the mean of the specified variable in the school-grade-subject-year cell. In Columns 1, 2, and 4, this variable is the end-of-year test score. In Column 3, it is the fitted value from a regression of end-of-year scores on parental characteristics taken from tax data (Panel A) or on parental education indicators (Panel B). In Columns 1-3, teachers observed only in a single year are excluded from the school-grade-subject-year mean predicted VA, and their students are excluded from the dependent variable. In Column 4, these teachers are assigned predicted VA of zero and are included, and their students are included in the dependent variable. See notes to CFR (2014a), Tables 4 and 5 for additional details about the specifications. Standard errors are clustered by school-cohort.

Table 2. Assessing the quasi-experiment via placebo tests

Dependent variable:	Δ prior year score	Δ predicted score given:	
		All VA model controls	Non-test VA model controls
	(1)	(2)	(3)
	Panel A: Excluding classrooms with missing teacher VA predictions		
Change in mean teacher predicted VA across cohorts	0.144	0.105	0.035
	(0.021)	(0.017)	(0.009)
# of school x grade x subject x year cells	79,466	78,186	79,466
	Panel B: Including classrooms with missing teacher VA predictions		
Change in mean teacher predicted VA across cohorts (all classrooms)	0.092	0.034	0.001
	(0.022)	(0.017)	(0.010)
# of school x grade x subject x year cells	90,701	88,949	90,203

Notes: Specifications in Panels A and B are identical to those in Table 1, Columns 2 and 4, respectively, but for changes in the dependent variable. In Column 1, this is the year-over-year change in mean prior year scores in the school-grade-subject-year cell. In Columns 2-3, it is the year-over-year change in mean predicted end of year scores in the cell. In Column 2, the predictions use all of the VA model controls, while in Column 3 only the non-test controls (indicators for race/ethnicity, gender, special education, free lunch status, limited english, and grade repetition; missing value indicators for each of these; and class- and school-year-level means of each) are used. Prediction coefficients are identified only from within-teacher variation. All specifications include school-year fixed effects, and standard errors are clustered by school-cohort.

Table 3. Adjusting the quasi-experiment for non-random assignment

Dependent variable:	Change in scores		Change in residual scores	Change in gain scores
	(1)	(2)	(3)	(4)
	Panel A: Without classrooms missing teacher VA prediction			
Change in mean teacher predicted VA across cohorts	1.030	0.933	0.931	0.889
	(0.021)	(0.015)	(0.014)	(0.015)
Change in mean prior year score		0.675		
		(0.004)		
# of school x grade x subject x year cells	79,466	79,466	78,186	79,466
	Panel B: Including all classrooms			
Change in mean teacher predicted VA across cohorts	0.904	0.860	0.894	0.832
	(0.022)	(0.017)	(0.015)	(0.017)
Change in mean prior year score		0.536		
		(0.009)		
# of school x grade x subject x year cells	91,221	90,701	88,949	90,692

Notes: Specifications in Panels A and B are identical to those in Table 1, Columns 2 and 4, respectively, but for changes noted here. In Column 3, the dependent variable is the the year-over-year change in mean residual scores, as defined in equation (2), in the school-grade-subject-year cell. In Column 4, it is the year-over-year change in mean gain scores, defined as the within-student difference between the end-of-year score and the prior-year score. Column 2 includes a control for the change in the mean score in the prior year. All estimates include school-year fixed effects, and standard errors are clustered at the school-cohort level.

Table 4. Association between teacher predicted VA and student characteristics

	Class level		School level
	Overall	Within school	
	(1)	(2)	(3)
Prior-year test score	0.063	0.028	0.394
	(0.005)	(0.002)	(0.047)
N	357,036	357,036	1,621
Free lunch	-0.022	-0.015	-0.106
	(0.003)	(0.001)	(0.031)
N	201,440	201,440	1,470
Minority student	-0.006	-0.009	0.035
	(0.003)	(0.001)	(0.035)
N	357,036	357,036	1,621
Predicted end-of-year	0.049	0.021	0.304
test score	(0.004)	(0.002)	(0.046)
N	349,322	349,322	1,621
Predicted college	0.0078	0.0018	0.064
enrollment	(0.0008)	(0.0003)	(0.008)
N	349,322	349,322	1,621

Notes: Each entry presents the coefficient from a separate regression of the indicated variable on the teacher's leave-one-out predicted VA score, rescaled into teacher-level standard deviation units (Columns 1-2), or on the school-level mean of this (Column 3). Column 2 includes school fixed effects. Regressions are weighted by the class or school size and standard errors are clustered at the school level.

Table 5. Observational analyses of teachers' long-run impacts

	# of classes	Teacher-year level regressions				
	(1)	(2)	(3)	(4)	(5)	(6)
		Panel A: CFR-II				
College at age 20 (%)	4,170,905	0.82	0.71	0.74		
		(0.07)	(0.06)	(0.09)		
College quality at age 20 ($)	4,167,571	298.6	265.8	266.2		
		(20.7)	(18.3)	(26.0)		
Earnings at age 28 ($)	650,965	349.8	285.6	309.0		
		(91.9)	(87.6)	(110.2)		
Variables used for within-teacher residualization of outcomes						
Baseline VA controls		X	X	X		
Parent chars.			X			
Twice lagged scores				X		
		Panel B: North Carolina replication				
Graduate high school (%)	2,318,646	0.34		0.27	0.24	0.22
		(0.04)		(0.05)	(0.04)	(0.04)
Plan college (%)	1,748,911	0.60		0.57	0.41	0.36
		(0.07)		(0.08)	(0.06)	(0.06)
Plan 4-year college (%)	1,748,876	1.35		1.45	0.87	0.73
		(0.09)		(0.11)	(0.08)	(0.08)
GPA (4 pt. scale)	1,191,964	0.022		0.009	0.018	0.016
		(0.002)		(0.002)	(0.002)	(0.002)
Class rank (100=top)	1,190,117	0.54		0.29	0.43	0.36
		(0.06)		(0.07)	(0.05)	(0.05)
Variables used for within-teacher residualization of outcomes						
Baseline VA controls		X		X	X	X
Twice lagged scores				X		
Controls in observational regression						
Baseline (classroom means)					X	X
Teacher means						X

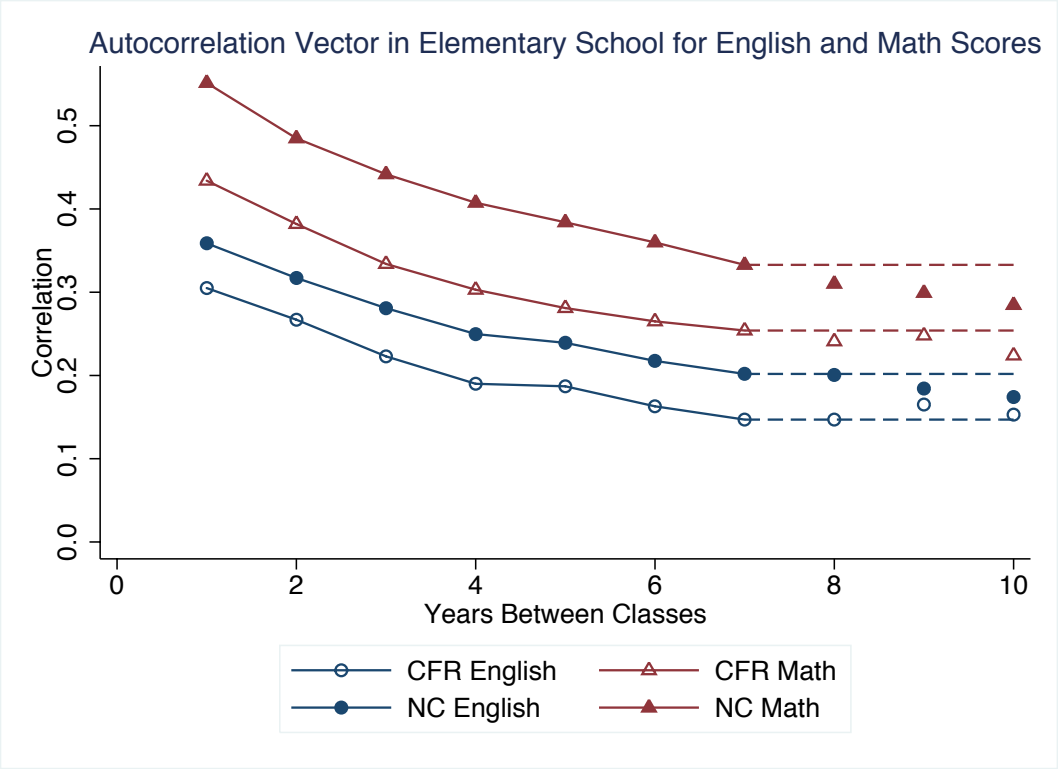
Notes: See notes to CFR-II, Table 2. Columns 2-4 report coefficients of regressions of residualized outcomes on teachers' predicted VA, varying the covariates used in residualizing the outcomes within teachers and controlling only for the subject to which the VA score pertains (math or reading) in the second stage regression. Columns 5 and 6 add classroom and teacher means of the VA covariates to the second stage regression. Standard errors are clustered at the school-cohort level. Column 1 shows the number of student observations used in the Column-2 regressions.

Table 6. Quasi-experimental estimates of effects on long-run outcomes

	Number of school x grade x subject x year cells	Quasi-experimental estimates	
		No controls	Prior score control
	(1)	(2)	(3)
	Panel A: CFR-II		
College at age 20 (%)	33,167	0.86	
		(0.23)	
College quality at age 20 ($)	33,167	197.6	
		(60.3)	
	Panel B: North Carolina		
Graduate HS (%)	50,508	0.38	0.26
		(0.17)	(0.17)
Plan college (%)	36,508	0.61	0.41
		(0.24)	(0.24)
Plan 4-year college (%)	36,508	0.45	0.09
		(0.27)	(0.26)
GPA (4 pt scale)	21,836	0.014	0.004
		(0.007)	(0.006)
Class rank	21,836	0.42	0.16
		(0.21)	(0.19)

Notes: Each entry in columns 2-3 represents a separate regression of the year-over-year change in school-grade-subject-year mean outcomes (indicated on left) on the change in mean predicted teacher VA. Each regression includes year fixed effects and is clustered at the school-cohort level. Column 3 also controls for the change in mean prior-year scores in the cohort. Following CFR-II, predicted VA is set to zero for teachers with missing predicted VA and for those who would otherwise be in the top 1% of the predicted VA distribution.

Appendix Figure 1
Reproduction of CFR-I, Figure 1A



Notes: See notes to CFR-I, Figure 1.

Appendix Table A1. Reproduction of CFR-I, Table 1 (Panel A only)
Summary statistics for sample used to estimate value-added model

	CFR-I, Table 1			North Carolina sample		
	Mean	SD	N	Mean	SD	N
	(1)	(2)	(3)	(4)	(5)	(6)
Class size (not student weighted)	27.3	5.6	391,487	22.2	5.0	357,036
No. of subject-years per student	5.6	3.0	1,367,051	4.5	1.7	1,607,198
Test score (SD)	0.2	0.9	7,639,288	0.0	1.0	7,215,581
Female	50.8%		7,639,288	49.7%		7,215,581
Age (years)	11.4	1.5	7,639,288	10.5	0.9	7,213,590
Free lunch elig.	79.6%		5,021,163	44.9%		3,926,246
Minority (Black/Hispanic)	71.6%		7,639,288	34.2%		7,215,581
English language learner	4.8%		7,639,288	8.5%		5,996,113
Special education	1.9%		7,639,288	2.3%		5,478,335
Repeating grade	1.7%		7,639,288	1.4%		7,215,581
Matched to parents in tax data	87.7%		7,639,288			

Notes: See notes to CFR-I, Table 1. In New York, free lunch eligibility is available only for 1999-2009. In North Carolina, it is available only for 1999-2006, and English language learner and special education information are available only 1997-2008.

Appendix Table A2. Reproduction 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	0.012	0.032
	(0.0003)	(0.0003)	(0.0002)	(0.0002)
	[0.305]	[0.434]	[0.359]	[0.551]
Lag 2	0.011	0.019	0.011	0.028
	(0.0003)	(0.0003)	(0.0002)	(0.0003)
	[0.267]	[0.382]	[0.317]	[0.485]
Lag 3	0.009	0.017	0.009	0.026
	(0.0003)	(0.0004)	(0.0002)	(0.0003)
	[0.223]	[0.334]	[0.281]	[0.442]
Lag 4	0.008	0.015	0.008	0.023
	(0.0004)	(0.0004)	(0.0002)	(0.0004)
	[0.190]	[0.303]	[0.250]	[0.407]
Lag 5	0.008	0.014	0.008	0.022
	(0.0004)	(0.0005)	(0.0002)	(0.0004)
	[0.187]	[0.281]	[0.239]	[0.384]
Lag 6	0.007	0.013	0.007	0.021
	(0.0004)	(0.0006)	(0.0003)	(0.0005)
	[0.163]	[0.265]	[0.218]	[0.360]
Lag 7	0.006	0.013	0.007	0.019
	(0.0005)	(0.0006)	(0.0003)	(0.0005)
	[0.147]	[0.254]	[0.202]	[0.333]
Lag 8	0.006	0.012	0.006	0.018
	(0.0006)	(0.0007)	(0.0003)	(0.0006)
	[0.147]	[0.241]	[0.201]	[0.310]
Lag 9	0.007	0.013	0.006	0.017
	(0.0007)	(0.0008)	(0.0003)	(0.0007)
	[0.165]	[0.248]	[0.184]	[0.299]
Lag 10	0.007	0.012	0.006	0.017
	(0.0008)	(0.0010)	(0.0004)	(0.0008)
	[0.153]	[0.224]	[0.174]	[0.285]
	Panel B: Within-Year Variance Components			
Total SD	0.537	0.517	0.561	0.544
Individual Level SD	0.506	0.473	0.542	0.495
Class+Teacher Level SD	0.117	0.166	0.144	0.225
Estimates of Teacher SD				
Lower Bound Based on Lag 1	0.113	0.149	0.110	0.180
Quadratic Estimate	0.124	0.163	0.118	0.192

Notes: See notes to CFR (2014a), Table 2. In Panel A, each entry includes the autocovariance, the standard error of that covariance (in parentheses), and the autocorrelation (in brackets) of average test score residuals across years, within teachers.

Appendix Table A3. Reproduction 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	1.021	0.009		0.022
	(0.004)	(0.001)		(0.002)
Parent Chars. Controls				
Observations	5,142,680	3,584,736		3,014,172

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 A4. Reproduction 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	1.097	1.030	0.994	0.008		0.202
across cohorts	(0.022)	(0.021)	(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	79,466	79,466	58,385	54,663		76,548

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

Appendix Table A5. Reproduction 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	1.174	0.936	1.100	1.081
across cohorts	(0.040)	(0.022)	(0.035)	(0.043)
Year Fixed Effects	X	X	X	X
Number of School x Grade x Subject x Year Cells	79,466	91,221	34,495	23,445
Pct. of Observations with Non-Imputed VA	100.0	72.6	94.4	100.0

Notes: See notes to CFR (2014a), Table 5. Panel B replicates CFR's estimates using the North Carolina sample.

Appendix Table A6. Reproduction of CFR (2014a), Table 6
Comparisons of Forecast Bias Across Value-Added Models

	CFR-I		North Carolina	
	Correlation with baseline VA estimates	Quasi-experimental estimate of bias (%)	Correlation with baseline VA estimates	Quasi-experimental estimate of bias (%)
	(1)	(2)	(3)	(4)
1. Baseline	1.000	2.58	1.000	-9.69
		(3.34)		(2.19)
2. Baseline, no teacher FE	0.979	2.23	0.981	-6.07
		(3.50)		(2.22)
3. Baseline, with teacher experience	0.989	6.66		
		(3.28)		
4. Prior test scores	0.962	3.82	0.976	-9.13
		(3.30)		(2.18)
5. Student's lagged scores in both subjects	0.868	4.83	0.955	-4.88
		(3.29)		(2.17)
6. Student's lagged score in same subj. only	0.787	10.25	0.923	-3.09
		(3.17)		(2.13)
7. Non-score controls	0.662	45.39	0.683	31.00
		(2.26)		(1.56)
8. No controls	0.409	65.58	0.522	46.41
		(3.73)		(1.32)

Notes: See notes to CFR-I, Table 6. CFR (2014a) do not provide code for the row 3 specification. Negative bias share coefficients in column 4 reflect estimated forecast coefficients above 1.

Appendix Table A7. 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						
Lagged test score	0.0122			0.0123			
	(0.0006)			(0.0006)			
Special educ. student		-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						
Lagged test score	0.0077						
	(0.0004)						
Special ed		0.0055					
		(0.0006)					
Minority (black/hispanic) student					-0.0028		
					(0.0012)		
School fraction minority							0.0054
							(0.0042)
N	5,142,680	5,142,680			5,142,680		5,142,680

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. The corresponding factor in the North Carolina sample (using CFR-I's calculation) is 1.36, and coefficients in Panel B are multiplied by that.

Appendix Table B1. Assessing potential mechanical contributions to the placebo test failure

		Excluding classrooms without VA predictions			Including all classrooms		
	Dependent variable	Δ Prior Year Score	Δ End-of-Year Score		Δ Prior Year Score	Δ End-of-Year Score	
			No controls	With control for Δ prior year score		No controls	With control for Δ prior year score
		(1)	(2)	(3)	(4)	(5)	(6)
1	Baseline	*0.14*	1.03	*0.93*	*0.09*	*0.90*	*0.86*
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
2	Cluster on school	*0.14*	1.03	*0.93*	*0.09*	*0.90*	*0.86*
		(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
3	IV setting VA of following teachers to zero	*0.08*	1.00	*0.95*	0.03	*0.87*	*0.87*
		(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
4	School-year-subject FEs	*0.12*	1.06	0.97	0.06	*0.91*	*0.89*
		(0.04)	(0.04)	(0.02)	(0.04)	(0.04)	(0.03)
5	School-year-subject FEs, IV	0.05	1.03	0.99	-0.02	*0.87*	*0.89*
		(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
6	Using leave-three-out teacher VA predictions	*0.17*	1.03	*0.92*	*0.12*	*0.91*	*0.85*
		(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
7	Leave-three-out, IV	*0.12*	1.01	*0.93*	*0.07*	*0.88*	*0.85*
		(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
8	Using leave-four-out teacher VA predictions	*0.16*	1.02	*0.91*	*0.13*	*0.90*	*0.84*
		(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)
9	Using leave-five-out teacher VA predictions	*0.15*	1.02	*0.91*	*0.13*	*0.89*	*0.83*
		(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)
10	Using leave-past-out teacher VA predictions	*0.14*	0.99	*0.89*	*0.12*	*0.88*	*0.82*
		(0.03)	(0.04)	(0.02)	(0.04)	(0.04)	(0.03)

Notes: Specifications in Row 1 correspond to Table 2, Column 1 (Cols. 1 and 4); Table 3, Column 1 (Cols. 2 and 5); and Table 3, Column 2 (Cols. 3 and 6). In each case, Columns 1-3 correspond to the Panel A specification in the earlier table, and Columns 4-6 to the Panel B specification. Successive rows modify the specification. In Rows 2-9, standard errors are clustered at the school level. In Row 3, the change in mean predicted teacher VA in the school-grade-subject-year cell is instrumented with a variable constructed similarly but with predicted VA set to zero for teachers who have ever previously taught the same cohorts. Row 4 presents OLS estimates with school-year-subject fixed effects, while row 5 reports IV estimates of the same specification using the non-following teacher instrument. In Rows 6-9, teacher VA predictions are constructed using only data from before t-2 (rows 6 and 7), t-3 (row 8), or t-4 (row 9). In Row 10, only data from after t is used. Row 7 applies the IV specification from Row 3 to the model from row 6, using leave-3-out VA predictions for non-follower teachers. Italicized coefficients are significantly different from the null hypothesis (zero in Columns 1 and 4; one in Columns 2, 3, 5, and 6).

Appendix Table B2. Assessing sensitivity of results to the imputation model

	Excluding classrooms missing teacher VA predictions	Including all classrooms, assigning to teachers with missing VA predictions:			
		Grand mean	School mean	Missing mean	Missing mean at school
	(1)	(2)	(3)	(4)	(5)
	Panel A: Quasi-experimental models without controls				
Change in mean teacher predicted VA	1.030	0.904	0.915	0.933	0.911
	(0.021)	(0.022)	(0.022)	(0.022)	(0.021)
	Panel B: Models for change in prior-year scores				
Change in mean teacher predicted VA	0.144	0.092	0.134	0.084	0.128
	(0.021)	(0.022)	(0.023)	(0.023)	(0.022)
	Panel C: Models for change in end-of-year scores, with controls for change in prior-year scores				
Change in mean teacher predicted VA	0.933	0.860	0.850	0.892	0.847
	(0.015)	(0.017)	(0.017)	(0.017)	(0.017)
Change in mean student prior year score	0.675	0.536	0.535	0.536	0.535
	(0.004)	(0.009)	(0.009)	(0.009)	(0.009)

Notes: Specifications in column 1, panels A-C are identical to those in Table 1, Column 2; Table 2, Column 1; and Table 3, Column 2, respectively. Successive columns include all classrooms in the dependent and independent variables, varying the VA prediction assigned to teachers who are excluded in column 1. In column 2, these teachers are assigned the grand mean of zero. In Column 3, the prediction is based on the shrunken leave-two-out mean at the same school. In Column 4, it uses the shrunken leave-two-out mean among all teachers with missing VA predictions. In column 5, it uses the shrunken leave-two-out mean among all teachers at the school with missing VA predictions. All specifications include school-year fixed effects. N=79,466 school-grade-subject-year cells in Column 1; 91,221 in Columns 2-5 in Panel A; and 90,701 in Columns 2-5, Panels B-C.

Appendix Table B3. Robustness of CFR-I, Table 5's robustness results
Quasi-Experimental Estimates of Forecast Bias: Robustness Checks

	Teacher Exit Only		Full Sample		<25% Imputed VA		0% Imputed VA	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Quasi-experimental models without controls							
Change in mean teacher predicted VA	1.174	1.080	0.936	0.904	1.100	0.965	1.081	0.918
	(0.040)	(0.044)	(0.022)	(0.022)	(0.035)	(0.040)	(0.043)	(0.051)
Year fixed effects	X		X		X		X	
School-year fixed effects		X		X		X		X
Number of School x Grade x Subject x Year Cells	79,466	79,330	91,221	91,221	34,495	34,495	23,445	23,445
	Panel B: Models for change in prior-year scores							
Change in mean teacher predicted VA	0.296	0.226	0.175	0.093	0.199	0.064	0.177	0.033
	(0.039)	(0.043)	(0.023)	(0.022)	(0.033)	(0.038)	(0.040)	(0.047)
	Panel C: Models for change in end-of-year scores, with controls for change in prior-year scores							
Change in mean teacher predicted VA	0.981	0.928	0.853	0.859	0.978	0.926	0.973	0.899
	(0.030)	(0.029)	(0.019)	(0.017)	(0.028)	(0.031)	(0.035)	(0.041)
Change in mean student prior year score	0.650	0.675	0.497	0.537	0.611	0.608	0.610	0.583
	(0.004)	(0.005)	(0.009)	(0.009)	(0.006)	(0.007)	(0.007)	(0.009)

Notes: See notes to CFR (2014a), Table 5. Columns 1, 3, 5, and 7 in Panel A reproduce results from that table. Even-numbered columns add school-year fixed effects. Panel B changes the dependent variable, while Panel C adds a control for the change in the prior-year score.

Appendix Table B4. Revisiting CFR (2015)'s simulations of missing VA and imputations

	Ideal Data (No Missing Values)	Exclude Obs with Missing Data	Impute 0s for Missing Data
	Dep. Var.: Change in Mean Score Across Cohorts		
	(1)	(2)	(3)
	Panel A: CFR (2015) Simulation: ICC = 0.35, VA known w/ certainty		
Change in Mean VA across Cohorts	0.989	0.972	
(dropping missing values)	(0.0248)	(0.0243)	
Change in Mean VA across Cohorts			0.879
(assigning zero if VA missing)			(0.0264)
Pct. Of Obs With Non-Imputed VA	100.0	100.0	80.0
Pct. Of Obs Excluded	0.0	20.0	0.0
	Panel B: ICC = 0.2, VA known w/ certainty		
Change in Mean VA across Cohorts	0.992	0.976	
(dropping missing values)	(0.0224)	(0.0226)	
Change in Mean VA across Cohorts			0.933
(assigning zero if VA missing)			(0.0245)
	Panel C: ICC = 0.35, VA predicted based on other years		
Change in Mean Predicted VA across Cohorts	0.863	0.912	
(dropping missing values)	(0.0273)	(0.0273)	
Change in Mean Predicted VA across Cohorts			0.825
(using prediction of 0 if no other data)			(0.0295)
	Panel D: ICC = 0.2, VA predicted based on other years		
Change in Mean Predicted VA across Cohorts	0.928	0.948	
(dropping missing values)	(0.0255)	(0.0260)	
Change in Mean Predicted VA across Cohorts			0.910
(using prediction of 0 if no other data)			(0.0280)
	Panel E: ICC = 0, VA predicted based on other years		
Change in Mean Predicted VA across Cohorts	0.992	0.987	
(dropping missing values)	(0.0235)	(0.0246)	
Change in Mean Predicted VA across Cohorts			0.999
(using prediction of 0 if no other data)			(0.0263)

Notes: See CFR (2015a), Table 2, and accompanying code in Appendix C. Panels B-E modify this code to change the correlation between VA scores of teachers at the same school (Panels B, D, and E) and to incorporate prediction of VA scores based on incomplete data as in CFR-I (Panels C, D, and E).