Title
Measuring the Impacts of Teachers: Comment

Journal
American Economic Review, 107(6)

ISSN
0002-8282

Author
Rothstein, Jesse

Publication Date
2017-06-01

DOI
10.1257/aer.20141440

Peer reviewed

American Economic Review 2017, 107(6): 1656–1684
https://doi.org/10.1257/aer.20141440

Measuring the Impacts of Teachers: Comment[†]

By Jesse Rothstein*

Chetty, Friedman, and Rockoff (2014a, b) 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 percent as large, in variance terms, as teachers' causal effects. Long-run results are sensitive to controls and cannot support strong conclusions. (*JEL* H75, I21, J45)

This comment 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 (VA) models of teacher effectiveness.[1]

The first paper, Chetty, Friedman, and Rockoff (2014a)—henceforth, CFR-I—attempts to measure bias in VA scores, interpreted as estimates of teachers' causal 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, Friedman, and Rockoff (2014b)—henceforth, 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

*Goldman School of Public Policy and Department of Economics, University of California, Berkeley, 2607 Hearst Avenue #7320, Berkeley, CA, 94720 (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.

[†]Go to https://doi.org/10.1257/aer.20141440 to visit the article page for additional materials and author disclosure statement.

[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.

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. Panel A of Figure 1 reproduces CFR-I's Figure 4, panel A, 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 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 estimates *forecast bias* (which I define more carefully below) as 1 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 (standard error 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 predetermined student characteristics that are predictive of outcomes. Figure 2 illustrates this. It is identical to panel B of Figure 1, 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 panel B of Figure 1, it is significantly and substantively greater than zero. CFR (2015a) have confirmed this result in the New York data, as have Bacher-Hicks, Kane, and Staiger (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 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 1.

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

[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, Kane, and Staiger'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 IIIB and the online Appendix.

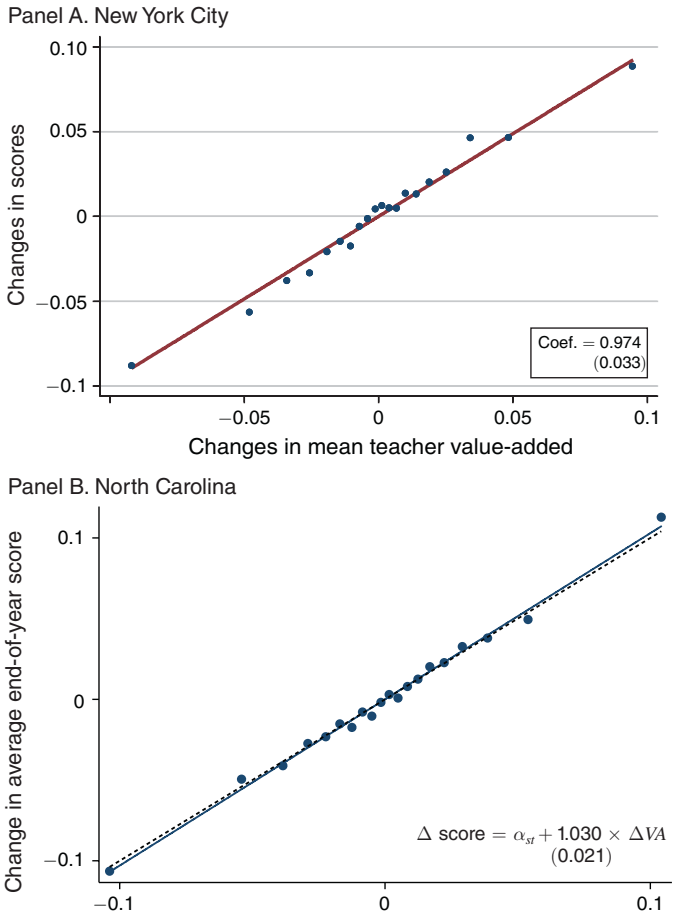

FIGURE 1. BIN SCATTERPLOT OF CHANGE IN AVERAGE TEACHER PREDICTED VA AND CHANGE IN AVERAGE END-OF-YEAR SCORE

Notes: Panel A is taken from CFR-I, Figure 4, panel A, 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 scatterplot 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 20 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 microdata 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 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 1. 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.

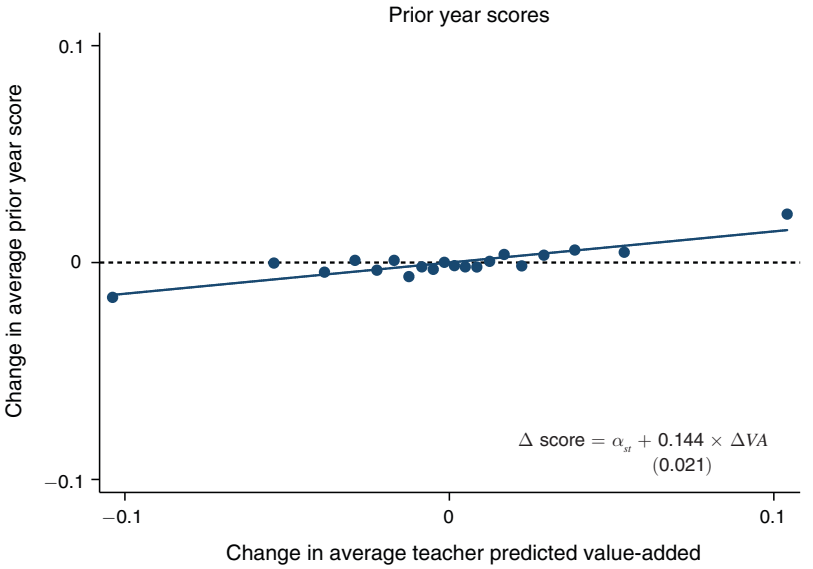

FIGURE 2. BIN SCATTERPLOT OF CHANGE IN AVERAGE TEACHER PREDICTED VA AND CHANGE IN AVERAGE PRIOR YEAR SCORE

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.

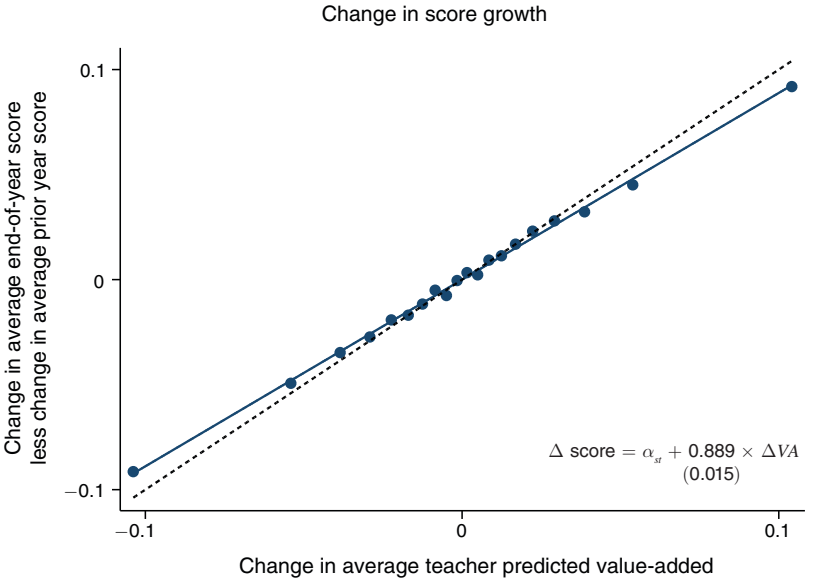

FIGURE 3. BIN SCATTERPLOT OF CHANGE IN AVERAGE TEACHER PREDICTED VA AND CHANGE IN AVERAGE GAIN SCORE

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.

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 nonrandom 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 nonzero—not as large as might 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 percent 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 percent or as large as 100 percent of the variance of teachers' true effects. Horváth (2015) estimates the correlation to be -0.3; if so, my estimates imply that the variance of the bias is nearly 35 percent 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 IV.

CFR-II presents both cross-sectional and quasi-experimental estimates of the association between teachers' VA scores and their impacts on long-run earnings.

[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. Rothstein (2017) explores robustness to alternative imputations, resting on different assumptions.

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

I show that the cross-sectional estimates, which do not control even for observed differences in teachers' students, rely on quite restrictive assumptions. 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 percent, 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, Kane, and Staiger 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 (i) VA scores are meaningfully, but not overwhelmingly, biased by student sorting, with forecast bias around 15 percent and (under reasonable assumptions) actual bias 10–35 percent as large, in variance terms, as teachers' causal effects, and (ii) teachers' VA scores are less informative than is implied by CFR-II's results, and perhaps completely uninformative, about the teachers' long-run impacts.

I. 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.

A. *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 factors above 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]

[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).

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

CFR-I refers 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.

B. *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 2009, 2010; Guarino, Reckase, and Wooldridge 2015). In the notation above, $\overline{A}_{jt}$ 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}.

[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.

[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 $\overline{A}_{jt} - \mu_{jt}$ is mean independent across years within teachers—that any nonrandomness in student assignments in any year is not persistent across years.

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 nonrandom 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 specializations—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 nonrandom 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 refers 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 defines *forecast bias* as $B \equiv 1 - \lambda$, where

$$\lambda \equiv \frac{\operatorname{cov}(\mu_j, \hat{\mu}_{jt})}{V(\hat{\mu}_{jt})} = \frac{V(\mu_j) + \operatorname{cov}(\mu_j, b_j)}{V(\mu_j) + V(b_j) + 2\operatorname{cov}(\mu_j, b_j)}. \tag{4}$$

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 $\operatorname{cov}(\mu_j, b_j) < 0$ then λ can equal or exceed 1 even when $V(b_j) > 0$. The available evidence suggests this is empirically relevant: Horváth (2015) estimates $\operatorname{corr}(\mu_j, b_j) = -0.3$ for North Carolina teachers, while Angrist et al. (forthcoming) estimate a correlation of -0.35 (with a large standard error) between schools' causal effects and the bias in school-level VA scores in Boston.

Rothstein (2009; see also Guarino, Reckase, and Wooldridge 2015) 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 $\operatorname{corr}(\mu_j, b_j) = 0$, he concludes that the plausible range for λ is roughly 0.6 to 1, corresponding to $V(b_j)/V(\mu_j)$ between 0 and $2/3$. If the correlation is instead -0.3, the upper bound of the variance ratio is about 0.75.

[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 nonzero 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.

C. *The Teacher-Switching Quasi-Experiment*

CFR-I builds 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 generalizes the experimental test to a nonexperimental 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 nonexperimental bias in mean outcomes. To abstract from this, CFR-I aggregates to the school (s)–grade (g)–subject (m)–year (t) level and consider changes in the *average* predicted VA of the teaching staff.[12] Their primary analyses regress the year-over-year change in mean student scores, $\Delta A^{*}_{sgmt} \equiv \overline{A}^{*}_{sgmt} - \overline{A}^{*}_{sgm,\,t-1}$, 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 excludes a nonrandom subset of classrooms from their aggregates. This biases the quasi-experimental coefficient toward the observational regression of $\overline{A}_{jt}$ on $\hat{\mu}_{jt}$, which necessarily—by virtue of the Empirical Bayes shrinkage used to construct $\hat{\mu}_{jt}$—has a coefficient of 1 regardless of the presence or absence of forecast or teacher-level bias.

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

[12] For their quasi-experimental analyses, CFR-I uses 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.

D. *Assessing the Quasi-Experiment*

The regression of ΔA^{*}_{sgmt} on ΔQ_{sgmt} identifies λ under CFR-I's Assumption 3 (henceforth, 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—with later cohorts more advantaged than earlier cohorts—are able to attract teachers who have higher (measured) VA than those whom 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 presents 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 online Appendix considers and rules out several potential mechanical explanations, proposed by CFR (2015a, 2014d) and Bacher-Hicks, Kane, and Staiger (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. In particular, the failure of the placebo test is robust to specifications that "isolate[...] sources of variation in teacher VA that are not

[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 Rothstein (2017).

spuriously correlated with prior test scores," as proposed by CFR (2015a). 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 nonrandom 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.[15]

II. 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 one-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 subdistricts. Just over 25 percent of students in North Carolina are Black and under 15 percent are Hispanic, with the remainder overwhelmingly white; in New York, about 30 percent are Black, 40 percent are Hispanic, 15 percent are Asian, and only 15 percent 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.[16] 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.

[15] Rothstein (2017) further explores the inclusion of missing classrooms in the sample, varying the strategy for assigning VA predictions to the missing teachers and restricting the sample to school-grade-year cells with no missing data, as suggested by CFR (2015a).

[16] 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 retake the tests, I use only the score from the first administration.

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.[17] This ensures that student achievement data are 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-grade mean prior scores; and class size.[18] For long-run outcomes, CFR-II draws 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 percent 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 versus 7.6 million in CFR-I's sample). I have nonmissing 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 nonmissing ΔQ_{sgmt}. I have 79,466 such cells, as compared with 59,770 in CFR-I.

III. The Teacher-Switching Quasi-Experiment: Reproduction and Assessment

A. *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.[19] Column 2 repeats the specification with school-year fixed effects.

[17] I use a less restrictive threshold for a valid assignment than in past work (e.g., Clotfelter, Ladd, and Vigdor 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 online Appendix.

[18] 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.

[19] 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 classrooms with teachers not seen in other years are omitted from both dependent and independent variables.

Table 1—Reproduction of CFR (2014a) Teacher Switching Quasi-Experimental Estimates of Forecast Bias

		Δ score		Δ score (predicted)	Δ score (all students)
		(1)	(2)	(3)	(4)
	Source:	T4C1	T4C2	T4C4	T5C2
Panel A. CFR (2014a)					
Change in mean teacher predicted VA across cohorts		0.974 (0.033)	0.957 (0.034)	0.004 (0.005)	
Change in mean teacher predicted VA across cohorts (with zeros)					0.877 (0.026)
Year fixed effects		X			X
School × year fixed effects			X	X	
Grades		4 to 8	4 to 8	4 to 8	4 to 8
Number of school × grade × subject 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 (0.022)	1.030 (0.021)	0.008 (0.011)	
Change in mean teacher predicted VA across cohorts (with zeros)					0.936 (0.022)
Year fixed effects		X			X
School × year fixed effects			X	X	
Grades		3 to 5	3 to 5	3 to 5	3 to 5
Number of school × grade × subject × 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.

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 (1.097) is somewhat larger than CFR-I's (0.974), and significantly greater than 1, but when I add school-year fixed effects in column 2, the coefficient (1.030) is much smaller and, like CFR-I's (0.957), indistinguishable from the null hypothesis. This is the specification illustrated in Figure 1.

CFR-I reports 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.[20] In both samples, the year-on-year change in mean predicted VA is uncorrelated with the change in mean predicted scores, with an estimated coefficient of 0.008 in North Carolina and 0.004 in New York.

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 $\hat{\lambda} = 0.877$ in New York and $\hat{\lambda} = 0.936$ in North Carolina. I discuss this result in more depth in the next subsection.

The online Appendix presents reproduction estimates for most of CFR-I's other analyses. Results are generally quite similar in North Carolina as in CFR-I's 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 (Table A2). In both math and English, the autocorrelation of teacher VA across years is higher in the North Carolina data (Table A2 and Figure A1 in the Appendix), 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.[21]

B. *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 fifth-grade teachers at school s between years $t-1$ and t, the dependent variable is the change in average fourth-grade scores across the same two cohorts (i.e., from $t-2$ to $t-1$).

[20] 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.

[21] Bacher-Hicks, Kane, and Staiger (2014) find that teacher VA is significantly *lower* in high minority share schools in Los Angeles.

Table 2—Assessing The Quasi-Experiment via Placebo Tests

	Δ prior year score (1)	Δ predicted score given: All VA model controls (2)	Δ predicted score given: Non-test VA model controls (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)
Number of school × grade × subject × 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)
Number of school × grade × subject × 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.

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 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, Kane, and Staiger (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 $\Delta \overline{A}^{*}_{sgmt} = \Delta \overline{A}_{sgmt} + \Delta \overline{X}_{sgmt}\hat{\beta}$, 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, e)

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.[22]

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.[23] This conclusively establishes that the placebo result cannot be attributed to the mechanical explanations proposed by CFR (2015a). [24]

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

[22] Note that either dynamic would likely invalidate not just the placebo test but also CFR-I's quasi-experimental research design itself (Rothstein 2017).

[23] The coefficient, 0.035 (SE 0.009), 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, Elder, and Taber (2005).

[24] The online Appendix explores this issue further. While there is some evidence that "teacher followers" contribute to the effect, the results are generally quite stable. See Table A8.

drops the teacher from the average Q_{sgmt} and drop the teacher's students from the average $\overline{A}_{sgmt}$.

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 online 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) $\overline{A}_{sgmt}$. 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 IIIA that CFR-I presents one specification that includes these teachers, assigning them predicted VA scores equal to the grand mean.[25] 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 assigns 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.[26]

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.[27]

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 constructs its 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 are 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

[25] 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.

[26] 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 includes these teachers, with predictions set equal to the grand mean. The issue here concerns only those teachers observed in one year but not the other. CFR do not explain the differential treatment.

[27] Other specifications, not reported here, indicate that the significant coefficients in panel B are—in contrast to the panel A results—not entirely robust.

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 this issue is important in practice. The intraclass 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. Rothstein (2017) 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.

Finally, it is important to note that excluding teachers with missing VA, 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, it would be best 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.

C. *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.[28] 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 nonmissing 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.[29] This reduces the $\hat{\lambda}$ coefficient to 0.933 (0.015).

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 $\Delta \overline{A}_{sgmt}$ rather than $\Delta \overline{A}^{*}_{sgmt}$). 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

[28] Note that the bias may be larger than the coefficients in Table 2, column 2 if unobservables change with observables—see footnote 23.

[29] CFR-I presents 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.

Table 3—Adjusting the Quasi-Experiment for Nonrandom Assignment

	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 (0.021)	0.933 (0.015)	0.931 (0.014)	0.889 (0.015)
Change in mean prior year score		0.675 (0.004)		
Number of school × grade × subject × 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 (0.022)	0.860 (0.017)	0.894 (0.015)	0.832 (0.017)
Change in mean prior year score		0.536 (0.009)		
Number of school × grade × subject × 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 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.

3. Note also that each of the methods for controlling for pretreatment 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 IIIB, 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 shows 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.[30]

The online Appendix (Table A8) 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. The $\hat{\lambda}$ coefficient is near 1 when the selected sample is used without adjustments 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

[30] 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.

observables. These results are not driven by any of the dynamics that CFR (2015a) point to as potential confounding factors. Rothstein (2017) presents additional specifications exploring alternative prediction strategies, other than assigning the grand mean, for the teachers excluded from CFR-I's main sample.

CFR-I presents 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 nonmissing 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 baseline specification using all cells and including only classrooms with nonmissing data. This appears to suggest that sample selection is a nonissue. But these estimates are quite imprecise, given the small sample. More important, CFR-I uses a different specification here, including only year effects where their preferred models include school-by-year fixed effects. Rothstein (2017) presents results of each specification.[31]

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 (Horváth 2015) implies $V(b_j)/V(\mu_j) \approx 0.35$. As I discuss in Section V, even the smaller estimate is large enough to produce a nontrivial misclassification rate (25 percent) 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.

IV. Long-Run Effects

The analysis thus far indicates that VA scores are moderately biased by student sorting, with forecast bias around 15 percent and teacher-level bias of 20–35 percent. 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 IVA with a discussion of the identification problem and CFR-II's observational strategy. I then present, in Subsection IVB, estimates of the long-run effects of North Carolina teachers, focusing on the sensitivity to the selection of controls and to the estimation strategy.

[31] 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.

A. *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:

$$\kappa \equiv \operatorname{cov}(m_j, \tau_j). \tag{5}$$

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 begins 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 $\overline{Y}_{jt}$ 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 nonpersistent sorting (ν_{jt}^Y) and random variation (e_{jt}^Y):

$$\overline{Y}_{jt} = \tau_j + b_j^Y + \nu_{jt}^Y + e_{jt}^Y. \tag{6}$$

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

$$\hat{\kappa} = \frac{\operatorname{cov}(\hat{m}_{jt}, \overline{Y}_{jt})}{V(\hat{m}_{jt})}. \tag{7}$$

Importantly, though CFR-II refers repeatedly to the inclusion of controls in this analysis (and CFR's Reply (2017) refers to "controls in our OLS regressions"), $\hat{\kappa}$ is always estimated via a bivariate regression; covariates are used only to construct the residual long-run outcome $\overline{Y}_{jt}$. 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 an OLS regression of $\overline{Y}_{jt}$ (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 $\overline{Y}_{jt}$. 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.[32] This is CFR-II's Assumption 1. As discussed above, the evidence suggests that it does not hold.

Second, $\overline{Y}_{jt} - \tau_j = b_j^Y + v_{jt}^Y + e_{jt}^Y$, 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 $\overline{Y}_{jt}$ in place of the teacher's causal effect τ_j would bias $\hat{\kappa}$.[33] This assumption is problematic as well. Where CFR-I argues that the bias in test score VA (b_j) was likely to be minimal, CFR-II finds affirmative evidence that teachers' estimated long-run impacts are biased—that is, that $V\left(b_j^Y\right) > 0$.[34] 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 online 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 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_j^Y with $\hat{\mu}_{jt}$.

Below, I show that $\hat{\kappa}$ is quite sensitive to the inclusion of controls for differences in observed student characteristics across teachers. This strongly suggests that $\hat{\kappa}$ is

[32] 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, $\tau_j - m_j\kappa$.

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

[34] 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 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).

Table 4— Association between Teacher Predicted VA and Student Characteristics

	Class level		
	Overall (1)	Within school (2)	School level (3)
Prior-year test score	0.063	0.028	0.394
	(0.005)	(0.002)	(0.047)
Observations	357,036	357,036	1,621
Free lunch	−0.022	−0.015	−0.106
	(0.003)	(0.001)	(0.031)
Observations	201,440	201,440	1,470
Minority student	−0.006	−0.009	0.035
	(0.003)	(0.001)	(0.035)
Observations	357,036	357,036	1,621
Predicted end-of-year test score	0.049	0.021	0.304
	(0.004)	(0.002)	(0.046)
Observations	349,322	349,322	1,621
Predicted college enrollment	0.0083	0.0023	0.065
	(0.0008)	(0.0003)	(0.008)
Observations	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.

biased when estimated without controls. But controls for student and family characteristics $\overline{X}_j$ change the estimand from κ to

$$\kappa_X \equiv \frac{\text{cov}\left(\mu_j,\, \tau_j \,|\, \overline{X}_j\right)}{V\left(\mu_j \,|\, \overline{X}_j\right)}. \tag{8}$$

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_j to $\overline{X}_j$ and imply that $\kappa_X \neq \kappa$.[35] Under this condition, it is exceedingly unlikely for $\text{cov}\left(b_j^Y, \mu_j\right) = 0$, 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 presents quasi-experimental analyses of teachers' long-run impacts analogous to those used to estimate forecast bias. I show below that these are as sensitive to the inclusion of controls for observables as are the corresponding short-run quasi-experimental estimates.

[35] 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 $\kappa_X = \kappa$.

B. *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 observes 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 points), on high school GPAs (0.022 GPA points), and on class rank (0.54 percentage points). All are highly statistically significant.

Columns 3 and 4 vary the controls used in estimating long-run VA $\overline{Y}_{jt}$, continuing to estimate (7) without controls. In column 2, the residualization uses just the covariates from the test score VA model. In column 3, CFR-II adds 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 percent, suggesting that bias in $\overline{Y}_{jt}$ 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 $\overline{Y}_{jt}$, but add controls to the second-stage regression of $\overline{Y}_{jt}$ 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 percent.[36]

[36] 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).

Table 5—Observational Analyses of Teachers' Long-Run impacts

	Number of classes	Teacher-year level regressions				
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. CFR-II						
College at age 20 (percent)	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 characteristics			X			
Twice lagged scores				X		
Panel B. North Carolina replication						
Graduate high school (percent)	2,318,646	0.34		0.27	0.24	0.22
		(0.04)		(0.05)	(0.04)	(0.04)
Plan college (percent)	1,748,911	0.60		0.57	0.41	0.36
		(0.07)		(0.08)	(0.06)	(0.06)
Plan 4-year college (percent)	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.

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 $\overline{X}_{jt}$ 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, $\Delta \overline{Y}^{*}_{sgmt}$, as examined by CFR-II. In their preferred

TABLE 6—QUASI-EXPERIMENTAL ESTIMATES OF EFFECTS ON LONG-RUN OUTCOMES

	Number of school × grade × subject × year cells	Quasi-experimental estimates	
		No controls	Prior score control
	(1)	(2)	(3)
Panel A. CFR-II			
College at age 20 (percent)	33,167	0.86	
		(0.23)	
College quality at age 20 ($)	33,167	197.6	
		(60.3)	
Panel B. North Carolina			
Graduate HS (percent)	50,508	0.38	0.26
		(0.17)	(0.17)
Plan college (percent)	36,508	0.61	0.41
		(0.24)	(0.24)
Plan 4-year college (percent)	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 percent of the predicted VA distribution.

specifications, and in contrast to CFR-I, CFR-II includes 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 percent), 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.

V. 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 1.

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 percent of the variance of teachers' causal effects. An estimate of $\hat{\lambda} = 0.9$ would correspond to a variance ratio of 11 percent. These are roughly in the middle of the range that Rothstein's (2009, 2010) simulations established as consistent with the data.[37] 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 Horváth 2015; Angrist et al. forthcoming). With a correlation of -0.3, teacher-level bias is 24 percent with $\lambda = 0.9$ and 32 percent 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 $\text{corr}(\mu_j, b_j) = 0$, 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.[38] If $\text{corr}(\mu_j, b_j) = -0.3$, 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 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 IV 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).

[37] CFR-I's VA model is most similar to Rothstein's (2010) VAM2. A variance ratio of 11 percent 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 percent is quite close to that in panel C.

[38] 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 one-half of teachers identified as in the bottom decile will have true μ_js outside of it. Misclassification rates are of course identical for teachers apparently in the top decile.

But these results turn out to depend on implausible assumptions. CFR-II's controls for student observables were implemented in a nonstandard 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 reopen 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, Kane, and Staiger (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 online Appendix. Rothstein (2017) presents additional specifications and robustness analyses.

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- versus low-VA teachers, which in the most credible estimates are not distinguishable from zero.

REFERENCES

Adler, Moshe. 2013. "Findings vs. Interpretation in 'The Long-Term Impacts of Teachers' by Chetty et al." *Education Policy Analysis Archives* 21 (10). http://epaa.asu.edu/ojs/article/view/1264.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113 (1): 151–84.

American Institutes for Research. 2015. "2013–14 Growth Model for Educator Evaluation: Technical Report." Washington, DC: American Institutes for Research. https://www.engageny.org/file/122791/download/2013-14-technical-report-for-growth-measures.pdf?token=59_KfVHr (accessed September 25, 2015).

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. Forthcoming. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics.*

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. 2014. "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles." National Bureau of Economic Research Working Paper 20657.

Ballou, Dale. 2012. "Review of 'The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.'" National Education Policy Center. http://nepc.colorado.edu/thinktank/review-long-term-impacts (accessed August 3, 2015).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2012. "Great Teaching." *Education Next* 12 (3).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–79.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014c. "Notes on Imputations and Controls for Observables." Unpublished.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014d. "Prior Test Scores Do Not Provide Valid Placebo Tests of Teacher Switching Research Designs." http://obs.rc.fas.harvard.edu/chetty/va_prior_score.pdf (accessed October 13, 2014).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014e. "Response to Rothstein (2014) on 'Revisiting the Impacts of Teachers.'" http://obs.rc.fas.harvard.edu/chetty/Rothstein_response.pdf (accessed October 13, 2014).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014f. "Stata Code for Implementing Teaching-Staff Validation Technique." http://obs.rc.fas.harvard.edu/chetty/cfr_analysis_code.zip (accessed July 21, 2014).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2015a. "Measuring the Impacts of Teachers: Response to Rothstein (2014)." Unpublished, July. http://obs.rc.fas.harvard.edu/chetty/va_response.pdf (accessed July 27, 2015).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2015b. "Measuring the Impacts of Teachers: Response to Rothstein (2014)." Unpublished, January.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2017. "Measuring the Impacts of Teachers: Reply." *American Economic Review* 107 (6): 1685–1717.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources* 41 (4): 778–820.

Deutsch, Jonah. 2013. "Proposing a Test of the Value-Added Model Using School Lotteries." Unpublished.

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge. 2015. "Can Value-Added Measures of Teacher Performance Be Trusted?" *Education Finance and Policy* 10 (1): 117–56.

Horváth, Hedwig. 2015. "Classroom Assignment Policies and Implications for Teacher Value-Added Estimation." Unpublished.

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

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14607.

Mansfield, Richard K. 2015. "Teacher Quality and Student Inequality." *Journal of Labor Economics* 33 (3): 751–88.

Rothstein, Jesse. 2009. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy* 4 (4): 537–71.

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

Rothstein, Jesse. 2014. "Revisiting the Impacts of Teachers." Unpublished.

Rothstein, Jesse. 2015. "Teacher Quality Policy When Supply Matters." *American Economic Review* 105 (1): 100–130.

Rothstein, Jesse. 2017. "Measuring the Impacts of Teachers: Comment: Dataset." *American Economic Review*. https://doi.org/10.1257/aer.20141440.

Rothstein, Jesse. 2017. "Supplement to 'Revisiting the Impacts of Teachers.'" http://eml.berkeley.edu/~jrothst/CFR/rothstein_CFR_supplement.pdf.

Rothstein, Jesse, and William J. Mathis. 2013. "Review of Two Culminating Reports from the MET Project." National Education Policy Center. http://nepc.colorado.edu/files/ttr-final-met-rothstein.pdf.

SAS Institute. 2015. Technical Documentation for 2015 TVAAS Analyses. Nashville, TN: Tennessee Department of Education. http://tn.gov/assets/entities/education/attachments/tvaas_technical_documentation_2015.pdf (accessed September 26, 2015).

Value-Added Research Center. Academic Growth over Time: Technical Report on the LAUSD School-Level AGT Model, Academic Year 2012–2013. Los Angeles: Los Angeles Unified School District. http://achieve.lausd.net/cms/lib08/CA01000043/Centricity/domain/414/documents/AGT%20Informative%20for%202010-2011.pdf (accessed September 26, 2016).

This article has been cited by:

1. Raj Chetty1, John N. Friedman2 and Jonah E. Rockoff3 1Stanford University, Landau Center 323, Stanford, CA 94305 (e-mail: chetty@stanford.edu) 2Brown University, Robinson Hall, Providence, RI 02912 (e-mail: john_friedman@brown.edu) 3Columbia University, Uris 603, New York, NY 10027 (e-mail: jonah.rockoff@columbia.edu) . 2017. Measuring the Impacts of Teachers: Reply. *American Economic Review* **107**:6, 1685-1717. [Abstract] [View PDF article] [PDF with links]