Measuring the Impacts of Teachers: Reply to Rothstein[1]

Raj Chetty, John Friedman, and Jonah Rockoff

January 2017

In two recent papers, Chetty, Friedman, and Rockoff (2014a,b) [CFR] measure teachers' effects on students' test scores and long-term outcomes. The first paper [hereafter, CFR-I] measures the degree of bias in teacher value-added (VA) estimates using a research design that exploits changes in teaching staff across years within schools, regressing changes in mean test scores across cohorts on changes in mean teacher VA. The second paper [hereafter, CFR-II] measures teachers' long-term impacts on students' earnings and other outcomes. CFR's main conclusions are that (1) teacher VA estimates exhibit little "forecast bias" – that is, they provide accurate predictions of teachers' *causal* effects on student achievement on average – and (2) teachers with high test-score VA also improve their students' long-term outcomes, such as college attendance rates and earnings.

Recent studies have replicated CFR's findings, which were based on data from New York City, using data from other school districts.[2] Jesse Rothstein (2016) successfully replicates CFR-I's results on bias in teacher VA using data from North Carolina and presents evidence on teachers' long-term impacts on students' outcomes in high school that is consistent with CFR-II's findings. Andrew Bacher-Hicks, Thomas Kane, and Douglas Staiger (2016) [hereafter BKS] replicate CFR-I's results using data from the Los Angeles Unified School District. These studies collectively demonstrate that value-added measures of teacher quality exhibit very consistent properties; as Rothstein (2016) puts it, "the empirical results are remarkably robust across disparate settings." Such replications are extremely valuable for evaluating the robustness of CFR's results and we are indebted to Rothstein and Bacher-Hicks, Kane, and Staiger for conducting these studies.

In addition to replicating CFR's findings, Rothstein (2016) raises three concerns about the validity of CFR's methods: the treatment of missing data, the validity of the quasi-experimental design, and the method of controlling for observables when estimating teachers' long-term effects. We are grateful that Jesse Rothstein has invested great effort in re-examining CFR's methodology, especially since our own interest in teacher value-added was sparked in part by Rothstein's (2010)

[1]We thank Gary Chamberlain, Lawrence Katz, and Michael Stepner for helpful comments and Augustin Bergeron, Nikolaus Hildebrand, and Benjamin Scuderi for research assistance.

[2]CFR-I and CFR-II did not name the school district, but it was identified in the *Vergara v. California* trial.

paper on this subject. We have learned a great deal from his work on this topic. However, we respectfully disagree with each of the points he makes. We show that Rothstein's methodological critiques are not valid by replicating each of his new empirical findings using simulated data in which none of CFR's identification assumptions are violated. We therefore conclude that Rothstein's findings are entirely consistent with CFR's methods and results. Our conclusions match those of BKS (2016), who also investigate the issues Rothstein raises in their data and conclude that they raise no concerns about the validity of CFR's analysis.

The rest of this introduction summarizes our responses to Rothstein's three main concerns and then situates this discussion in the context of the broader policy debate on value-added methods.[3]

Rothstein Concern #1 [Missing Data]: The exclusion of teachers with missing VA data leads CFR to underestimate the degree of forecast bias in VA measures.

Response: The most definitive way to evaluate whether missing data are a concern is to focus on the subset of schools where no data are missing. In this subsample, there is no evidence of forecast bias in New York (CFR-I, Table 5, Column 4), North Carolina (Rothstein 2016, Appendix Table A5, Column 4), or Los Angeles (BKS 2016, Table 3, Column 5). Rothstein acknowledges that these results "appear to suggest that sample selection is a non-issue," but goes on to argue that the estimates in the subsample with no missing data are "quite imprecise." The standard errors of the forecast bias estimates increase from about 2% to 4% when one restricts to the subsample with no missing data in all three datasets. We disagree with the view that an increase of 2 percentage points in the standard error makes an estimate of forecast bias – which could theoretically range from 0% to 100% – "quite imprecise." Notably, the estimate remains precise enough to reject Rothstein's own estimates using alternative methods of handling missing data with $p < 0.05$.

Instead of focusing on subsamples with no missing data, Rothstein imputes VA for teachers in cases where it is missing. We present a simple statistical derivation and a simulation showing that this imputation procedure generates downward-biased estimates of the effect of VA on test scores even when VA is missing at random, inflating the estimate of forecast bias. This is because Rothstein's preferred imputation introduces measurement error in changes in VA across cohorts when VA is correlated across teachers within school-grade cells, as is the case empirically.[4]

[3]In the interest of full disclosure, note that much of this paper is taken from response letters that we submitted to the *American Economic Review* during the review process of the CFR papers. In particular, two of Rothstein's concerns – on imputation and long-term controls – were raised and addressed during the referee process. Rothstein was one of the referees of the CFR papers – a fact that was disclosed during his testimony for the defense in *Vergara v. California*.

[4]Rothstein (2016, Section 3.2) acknowledges that this imputation procedure generates bias, but suggests that

We conclude that (a) from a theoretical perspective, Rothstein's estimates with imputed data are inconsistent and (b) from an empirical perspective, the results remain very similar in subsamples with no missing data, implying that CFR's treatment of missing data does not produce bias.

Rothstein Concern #2 [Prior Score "Placebo" Test]: The research design used by CFR to estimate forecast bias and long-term impacts is invalid because it fails a placebo test showing "effects" of changes in mean teacher VA on changes in prior test scores.

Response: These apparent "placebo effects" stem from a problem with the placebo test itself rather than a flaw in CFR's research design. Using stylized examples and simulations, we demonstrate that Rothstein's "placebo test" detects a correlation between prior test score changes and current teacher VA even when the CFR research design is valid. As a result, one cannot conclude from such a finding that the CFR design is invalid.

Prior test scores cannot be used to conduct placebo tests because VA is itself estimated from prior test scores. Intuitively, teachers will tend to have high VA estimates when their students happened to do well in prior years due to a positive shock. Regressing changes in prior test scores on changes in teacher VA effectively puts the same data on the left- and right-hand sides of the regression, which mechanically produces a positive coefficient even with a valid research design.[5] Thus, placebo tests using prior scores are not, as Rothstein suggests, akin to checking for balance on exogenous pre-determined variables in a randomized experiment, because the "treatment effect" in this setting (VA) is endogenously estimated from data on test scores.[6] For the same reason, Rothstein's estimates that control for prior test scores are inconsistent, as part of the effect of changes in teacher VA is incorrectly attributed to changes in prior test scores that are spuriously correlated with VA.

The results described above show that prior test scores are uninformative about the validity of the research design and hence do not raise concerns about CFR's analysis. Nevertheless, to address any concerns that the prior test score correlation may actually be driven by selection bias,

the bias is likely to be small. We present Monte Carlo simulations showing that, on the contrary, the bias due to imputation is of the same order of magnitude as the change in the estimate of forecast bias when imputed data are included in both New York and North Carolina.

[5]Rothstein (2016) suggests that such mechanical effects can be addressed by leaving out additional years of data when estimating teacher VA, but we show that this is not true when shocks to test scores are serially correlated, as is the case empirically.

[6]Rothstein finds small correlations between certain exogoneous characteristics (such as an indicator for low family income based on free-lunch eligibility) and changes in teacher VA in North Carolina, and shows that controlling for these factors reduces the forecast coefficient by 3.5 percentage points (Rothstein 2016, Table 2a, Column 3), from 93.3% to 89.8%. This result is not reproduced in New York: CFR-I conducted analogous tests for balance using exogenous characteristics (such as parent income), and found no such correlations (CFR-I, Table 4, Column 4).

we show that addressing some of the simplest mechanical effects eliminates the correlation between changes in lagged scores and current teacher VA but does not affect the estimate of forecast bias. In particular, simply excluding teachers who taught the same students in the previous year reduces the coefficient on prior test scores from 0.171 to a statistically insignificant 0.052 in the New York data. Excluding all *within-school* switchers – whose VA is estimated from historical data in the same school – yields a coefficient on prior scores of 0.031. None of these changes in the specification significantly changes the coefficient on current scores (i.e., the estimate of forecast bias). Both Rothstein (2016, Appendix Table A8, Row 5) and BKS (2016, Table 4, Column 6) obtain the same results – an insignificant coefficient close to 0 on prior scores and a coefficient close to 1 on current scores – when replicating these specifications in North Carolina and Los Angeles.

We conclude that (a) from a theoretical perspective, Rothstein's prior test score regression is not a valid placebo test and (b) from an empirical perspective, CFR's quasi-experimental estimates of forecast bias remain similar even when one isolates sources of variation in teacher VA that are not spuriously correlated with prior test scores.

Rothstein Concern #3 [Estimation of Long-Term Impacts]: The estimates of teachers' long-term impacts based on OLS regressions rely on strong assumptions that may not hold in practice and the quasi-experimental estimates may be upward-biased.

Response: Rothstein observes that CFR's method of controlling for observables, which uses within-teacher variation to identify the impacts of covariates, relies on strong assumptions. He proposes an alternative approach that uses between-teacher variation for identification. We demonstrate using simulations that Rothstein's approach yields inconsistent estimates under the assumptions of CFR-II's statistical model. Our OLS regression methodology provides an internally consistent approach to estimating teachers' long-term impacts under the identification assumptions stated in CFR-II, whereas Rothstein's proposed methodology does not.

Although Rothstein offers no definitive evidence that the identification assumptions made in CFR-II are violated, he is certainly correct to raise the possibility that the controls in our OLS regressions might not adequately account for omitted variable bias; indeed, we emphasized on page 2 of CFR-II that this is "a very strong assumption." This is why we returned to our quasi-experimental research design to obtain empirical estimates that do not rely on such assumptions (CFR-II, Section IV). We found that the quasi-experimental method yields very similar estimates of teachers' long-term impacts for the outcomes for which we have adequate precision (e.g., college attendance). Rothstein (2016, Table 6) replicates this finding but then (a) argues that these

quasi-experimental estimates are upward-biased because controlling for prior test scores reduces the estimated effects, and (b) interprets his estimates as providing "no strong basis for conclusions about the long-run effects of high- vs. low-VA teachers, which in the most credible estimates are not distinguishable from zero." We disagree with both of these conclusions. First, controlling for prior test scores in the quasi-experimental specifications for long-run outcomes yields invalid estimates for the same reason as discussed above in our response to Concern #2. Second, Rothstein's failure to find significant effects reflects the lack of statistical power in the North Carolina sample, not a lack of evidence regarding long-term effects in the New York data. In his quasi-experimental specifications, Rothstein can rule out neither zero effects, nor his own OLS estimates, nor effects (based on plans for college attendance) larger than those estimated in the New York data.

We conclude that (a) from a theoretical perspective, Rothstein's critiques of CFR's method of controlling for covariates in OLS regressions and of the quasi-experimental estimates are invalid and (b) from an empirical perspective, CFR-II's estimates of teachers' long-term effects are not inconsistent with the less precise quasi-experimental estimates reported by Rothstein.

In summary, each of Rothstein's three critiques of CFR's analysis is not valid: his preferred method of imputing teacher VA generates bias, his proposed prior score placebo test falsely rejects valid research designs, and his method of controlling for covariates yields inconsistent estimates of teachers' long-term effects. Excluding these incorrect results, Rothstein's analysis consists of (1) an exact replication of CFR-I's analysis showing that standard VA models exhibit little forecast bias in data from North Carolina and (2) new estimates of high-VA teachers' positive long-term effects using the methods of CFR-II. Thus, Rothstein's rigorous analysis ultimately increases our confidence in CFR's results. These conclusions echo those of BKS (2016), who write that "[we] found that teacher value-added estimates were valid predictors of student achievement. We also demonstrate that Rothstein's [prior score] test does not invalidate the CFR design and instead reflects a mechanical relationship, given that teacher value-added scores from prior years and baseline test scores can be based on the same data."

Taking a step back, it is helpful to situate this discussion about bias in VA estimates in the context of the broader debate about VA as a tool to evaluate teacher quality. CFR estimate that 95.7% of the difference in VA between teachers reflects causal impacts on student achievement (CFR-I, Table 4, Column 2). Rothstein (2016, Table 3a, Column 2) finds that controlling for lagged test score changes reduces this estimate to 93.3% in the North Carolina data. Imputing VA for teachers with missing data further reduces the estimate to 86.0% (Rothstein 2016, Table 3b,

Column 2). In short, all three quasi-experimental studies (including BKS' Los Angeles analysis) agree that at least 86% of the differences in VA between teachers reflect causal impacts on student achievement. Similarly, even Rothstein's preferred estimates of the long-term effects of teacher VA on high school graduation and planned college attendance rates – which are about 30% smaller than his estimates using CFR-II's specifications – would still imply very large gains from improving teacher VA.[7] In contrast with Rothstein's interpretation of these results, our assessment is that they show that VA measures exhibit remarkably consistent properties across three large, disparate datasets. From a policy perspective, resolving whether 86.0%, 93.3% or 95.7% of differences in VA reflect causal effects is not critical for deciding whether evaluating teachers based on their VA can produce substantial expected gains for students.

Rothstein argues that the magnitude of these differences is larger than suggested by the preceding discussion by turning to a different notion of bias, which CFR-I term "teacher-level" bias: the extent to which individual teachers' VA estimates differ in expectation from their true effects. As discussed in CFR-I (pages 2601-2602 and Online Appendix B), the relevant notion of bias depends upon the policy question one seeks to answer. If we are interested in calculating the expected gains from policies that select teachers based on their VA, as in CFR-II, then forecast bias is what matters: forecast bias of 10% reduces the expected gains from selecting teachers based on VA by 10%. If we are interested in the degree to which teachers may be misclassified based on VA estimates relative to their true effects, then teacher-level bias becomes relevant. CFR-I's research design was designed to estimate forecast bias and does not itself produce any estimates of teacher-level bias. As a result, the present debate does not have clear implications for the degree of teacher-level bias: even if forecast bias were smaller than estimated by CFR, teacher-level bias could potentially exceed the amount suggested by Rothstein. Rothstein makes specific structural assumptions to translate his estimate of forecast bias to predictions about teacher-level bias. Our view is that teacher-level bias should instead be estimated directly in future work using new quasi-experimental or experimental designs. This is why we focus on the degree of forecast bias – the parameter that CFR, Rothstein, and BKS actually estimate – to assess how much their results differ.

The remainder of this note is organized as follows. Section I addresses missing data on teacher VA. Section II addresses the prior test score "placebo test." Section III discusses the estimation of long-term impacts. Section IV concludes.

[7] We focus on Rothstein's results for high school graduation and planned college attendance because they are most similar to the outcomes studied by CFR-II. As we discuss in Section III.A, teachers' effects on the other outcomes Rothstein examines – high school GPA and class rank – need not mirror their effects on outcomes in adulthood.

I Missing Data on Teacher Value-Added

The first concern raised in Rothstein (2016, page 20) is that CFR's treatment of teachers with missing value-added leads them to understate the degree of forecast bias in VA estimates.

To help orient the reader, we first briefly review the quasi-experimental research design developed in CFR-I to estimate the degree of forecast bias in VA. The design exploits teacher turnover at the school-grade level for identification. For example, suppose a high-VA 4th grade teacher leaves school A to teach at another school in 1995. Because of this staff change, 4th graders in school A in 1995 will have lower VA teachers on average than the previous cohort of students in school A. If VA estimates have predictive content, we would expect 4th grade test scores for the 1995 cohort to be lower on average than the 1994 cohort. Building on this intuition, we estimate the amount of forecast bias by regressing changes in average test scores across consecutive cohorts of children within a school on changes in the mean value-added of the teaching staff. CFR-I estimate that a 1 unit change in mean VA across cohorts leads to a $\lambda \simeq 0.96$ unit increase in test scores. The estimated amount of forecast bias is therefore $B = 1 - \lambda = 4\%$ and is not statistically distinguishable from 0.

Our teacher switchers research design uses leave-two-year-out estimates of VA (i.e., VA estimates based on test score data excluding 1994 and 1995 in the example above). This two-year leave-out approach eliminates mechanical relationships between the change in VA and the change in test scores that can arise from noise in student test scores. Since some teachers teach for only one or two years before leaving the school district, we do not have VA estimates for 16% of the teacher observations in our New York data. This is the root of the missing data problem that this section discusses.

In our baseline analysis, we exclude teachers with missing VA from our sample, recognizing that this could lead to selection bias. As we noted in CFR-I (Section V.E), "restricting the sample to classrooms with non-missing VA estimates could lead to violations of our identification assumption because we do not use the entire school-grade-subject cohort for identification." Missing data on teacher VA can potentially create selection bias because of non-random assignment of students to classrooms within a school-grade-subject-year cell.[8] We addressed this concern about selection bias

[8]To take an extreme example, suppose teacher information is reported for only one of 5 classrooms in a given school-grade-subject cell. In this case, a comparison of mean scores across students in the two cohorts with non-missing teacher information is equivalent to comparing mean scores across a single classroom in two different years. This could violate the identifying assumption of CFR's quasi-experimental design if assignment to classrooms is non-random.

by restricting the sample to school-grade-subject-year cells with no missing teacher data – i.e., cells where the missing data problem does not arise – and we obtained very similar estimates in this subsample (Column 4 of Table 5, CFR-I). Rothstein and BKS replicate these findings in North Carolina and Los Angeles (Rothstein 2016, Appendix Table A5, Column 4; BKS 2016, Table 3, Column 5). Despite wide variation in the extent of missing data – 27% in North Carolina, 16% in New York, and 8% in Los Angeles – all three analyses using the restricted sample suggest negligible bias in VA.

Rothstein presents no argument for why estimates of forecast bias in the subset of schools with no missing data are invalid. Instead, he argues that an alternative approach – imputing a value of 0 for teachers with missing VA – is a better way to deal with missing data. He observes that one obtains larger estimates of forecast bias when one uses this approach (i.e., a lower estimate of the VA prediction coefficient λ). We had implemented this approach in our original paper, and like Rothstein, we show that imputing 0's for teachers with missing VA yields larger estimates of forecast bias (Cols 2-3, Table 5, CFR-I). In CFR-I, we explain that "this is to be expected because the imputation of teacher VA generates measurement error in average teaching quality, leading to attenuation bias. For example, if a highly effective teacher enters the district for only a single year, so that we are not able to calculate VA from other years of data, our imputation procedure will treat this teacher as being average, leading to measurement error in mean VA in the school-grade-subject cell."

This section recapitulates and expands upon the above points. We begin by reviewing the empirical results which show that both in our data and Rothstein's data, restricting the sample to school-grade-subject-year cells with no missing teacher data yields estimates that are very similar to our baseline results. We then present a simple derivation which shows that imputing a value of 0 to teachers with missing data yields attenuated estimates of λ when VA is correlated across teachers within a cell and thus overstates forecast bias B, just as one finds empirically.

I.A Empirical Estimates from Subsamples Without Missing Data

In any empirical application, the simplest and most direct method of evaluating the effects of missing data is to find a sample (e.g., a set of schools) where there is no missing data. We implement this approach by restricting the sample to school-grade-subject-year cells with no missing teacher VA data in both the current and preceding year. Restricting the sample at the school-grade-subject level does not generate selection bias because we exclude *entire cohorts* rather than individual

classrooms. Selection bias arises from potential non-random assignment of students to classrooms *within* school-grade-subject-year cells. When we exclude entire school-grade-subject-year cells from the analysis, we obtain a consistent estimate that is free from such selection bias, albeit a local average treatment effect that applies to schools with non-missing data.

Table 1 reproduces a set of specifications from CFR-I (Tables 4 and 5), Rothstein (2016, Appendix Tables A4 and A5), and BKS (Table 3) that conduct such an analysis. Column 1 shows that the baseline approach of using the entire sample and excluding teachers with missing VA data generates prediction coefficients of $\lambda \simeq 1$ in all three datasets. The estimate is 0.974 (s.e. = 0.033) in New York, 1.097 (s.e. = 0.022) in North Carolina, and 1.03 (s.e. = 0.044) in Los Angeles.[9]

Column 2 restricts the sample to observations to school-grade-subject-year cells in which there is no missing VA data in both the current and preceding year. In this subsample, which includes 30% of the school-grade-subject-year cells in the New York data, $\lambda = 0.990$ (s.e. = 0.045). Similarly, Rothstein finds that in the North Carolina data, when he restricts to cells with no missing data, the resulting coefficient is $\lambda = 1.081$ (s.e. = 0.043). For the non-missing subsample in Los Angeles, BKS (2016) report a coefficient of 0.973 (s.e. = 0.048). Rothstein acknowledges that these results "suggest that sample selection is a non-issue," but then dismisses these estimates as "quite imprecise," even though the standard errors differ from the baseline estimates by less than 0.02 and are sufficiently small to rule out the estimate he obtains using imputation for missing data.[10]

Of course, the specification in Column 2 of Table 1 only provides an estimate of forecast bias in the subsample without missing data. The schools without any missing data are very similar to the district as a whole on observables, suggesting that the degree of forecast bias in that sample is likely to be representative of forecast bias in general. Nevertheless, to assess the stability of the estimate across samples, in Column 3 of Table 1, we restrict the sample to school-grade-subject-year cells where VA is missing for less than 25% of the observations. This sample covers 65% of the school-grade-subject-year cells in the New York data. Following Rothstein's preferred procedure of imputing a value of 0 (the sample mean) for teachers with missing VA, we obtain an estimate of $\lambda = 0.952$ (s.e. = 0.032). We obtain $\lambda \simeq 1$ in this sample despite imputing 0 for missing data – which we show below yields biased estimates of λ – because VA is missing for only 6% of the

[9] Why the estimate in North Carolina is greater than one is unclear; an earlier version of Rothstein's paper (2014) reported a coefficient of 1.050 (s.e. = 0.023) from a regression with three percent fewer observations. However, Rothstein (2016) reports that adding school-year fixed effects, rather than year fixed effects, brings the estimates much closer to one.

[10] Rothstein also notes that these specifications include year fixed effects, rather than school-by-year fixed effects. In the New York data, specifications with year- and school-by-year fixed effects yield estimates that are not statistically distinguishable from one another.

observations in this subsample.[11]

The key lesson of Columns 1-3 of Table 1 is that the baseline estimates of λ are reproduced almost exactly in subsamples with little or no missing VA information (which cover two-thirds of the New York data). We therefore conclude that CFR-I's baseline estimates of forecast bias are not substantively biased by the exclusion of teachers with missing VA. We believe that any proposal for why missing data creates bias should first explain why the results do not change in subsamples with little or no missing data. Rothstein (2016) does not offer such an explanation; instead, he proposes an imputation procedure that yields different results, which we turn to next.

I.B Parametric Imputation of VA with Missing Data

An alternative method of dealing with missing data is to model the underlying data-generating process and parametrically impute a VA estimate for teachers with missing data. The simplest approach is to assume that the VA of teachers with missing data is drawn at random from a distribution with mean 0 (that is, with a mean equal to sample-wide average teacher quality). Importantly, this assumption implies that teacher VA is uncorrelated across teachers within the same school-grade-subject-year cell. Under this assumption, imputing $\hat{\mu}_{jt} = 0$ (the unconditional "grand" mean) for teachers with missing data yields unbiased estimates of forecast bias. Column 4 of Table 1 replicates the specification in Column 1, including teachers with missing VA data by assigning them a VA of 0. This yields a prediction coefficient of $\lambda = 0.877$ in the New York data and 0.936 in the North Carolina data – noticeably lower than the estimates obtained in the preceding columns.

Why does imputing 0 to teachers with missing VA reduce the estimate of λ? One problem is that the crucial assumption – that VA is uncorrelated across teachers within a school-grade-subject cell – does not hold in practice. In the New York data, the VA of those entering and leaving a given school-grade-subject cell has a correlation of approximately 0.2, consistent with well-established results in the literature that teacher quality varies systematically across schools because of sorting and other factors. That is, if teacher A (who is leaving the cell) was a low VA teacher, teacher B (who replaces her) tends to be a low VA teacher as well. Now suppose VA is

[11] While roughly half of the school-grade-year cells with any missing VA data in New York pass the "25% or less" restriction, this is true of only approximately 15% of cells with missing VA in North Carolina, so there is little to be gained in terms of generalizeability from this specification in Rothstein's sample. Imposing this restriction produces an estimate of $\lambda = 1.100$ (s.e. $= 0.035$) in the North Carolina data, similar to the sample with no missing data. BKS (2016) do not report estimates for the "25% or less missing" sample, presumably because their subsample with no missing VA already covers 63 percent of the Los Angeles data.

missing for teacher B. If we impute 0 to teacher B, we systematically overstate the change in VA relative to the truth. This exaggerates the change in VA across cohorts (the independent variable in our cross-cohort quasi-experimental regression specification) and therefore biases the estimate of λ toward 0. Conceptually, when the assumption that the VA of missing teachers is drawn at random is violated, 0 is no longer an unbiased forecast of the VA for an entering/exiting teacher with missing data.

Another way to see this point is to consider the experimental study by Kane and Staiger (2008), hereafter KS, where pairs of teachers within the same school-grade-subject-year cell were randomly assigned one of two classrooms. KS run a regression of the difference in class average end-of-year test scores on the difference in VA between the two teachers. Suppose that KS were missing a VA estimate for one of the two teachers in some of their randomization cells, and that these cells with missing data were selected at random. Rather than drop the cell – which is analogous to what we do above in Column 2 of Table 1 – Rothstein would suggest imputing VA of zero for the missing teacher. If VA is positively correlated across teachers within cells, as is true in practice, then imputing the unconditional mean (0) to teachers with missing data will overstate within-pair differences in VA. This will attenuate the prediction coefficient obtained from the experiment, leading to an upward-biased estimate of the degree of forecast bias.

Statistical Derivation. To formalize this logic, we present a derivation showing the bias due to imputation in our switchers design. Consider a simplified version of the statistical model in CFR-I in which there are many schools s, each with one grade and one subject. We have data on students' test scores for two years, $t \in \{1, 2\}$, and use changes in the teaching staff between years $t = 1$ and $t = 2$ to estimate forecast bias. To focus on the problem of missing data, we assume that if a teacher's VA is observed, it is observed without error.

Each school has three classrooms of equal size, taught by three different teachers. Teachers K and M are present in the school in both years 1 and 2. Teacher L is present in year 1 and is replaced by a new teacher Z in year 2. Let μ_{js} denote the VA of teacher j in school s, which we assume is fixed over time for simplicity (i.e., there is no drift in teacher quality). The change in average VA between years 1 and 2 in school s is $\Delta\bar{\mu}_s = \frac{\mu_{Zs} - \mu_{Ls}}{3}$. Let $Corr\left(\mu_{js}, \mu_{j's}\right) = \rho$ denote the correlation between the VA of any two teachers j and j' who teach in the same school. Denote by ΔQ_s the measured change in VA.

The mean test score of students in the class taught by teacher j in year t in school s is

$$\bar{A}_{jts} = \mu_j + \varepsilon_{jts},$$

where we ignore the effects of covariates for simplicity.

We begin with the case where all teachers' VA are observed. Let $\Delta\bar{A}_s$ denote the change in mean test scores for students between the two years in school s and $\Delta\bar{\mu}_s$ the change in mean VA. The quasi-experimental cross-cohort regression is

$$\Delta\bar{A}_s = \lambda\Delta Q_s + \nu_s,$$

where ν_s denotes the difference in average error terms within each school. To highlight the effects of missing data, we assume the quasi-experimental design is otherwise valid, so that $Cov\left(\Delta\bar{\mu}_s, \nu_s\right) = 0$. With no missing data, where $\Delta Q_s = \Delta\bar{\mu}_s$, it is clear that this regression would yield a consistent estimate of the true coefficient, $\lambda = 1$.

Now suppose we have no data on the VA of teacher Z, the entering teacher, and teacher M, one of the teachers who is present in both years. Following Rothstein's preferred approach, suppose we impute $\mu_Z = 0$ and $\mu_M = 0$ and run the regression

$$\Delta\bar{A}_s = \lambda_{imp}\Delta Q_s + \nu_s. \tag{1}$$

Since $\Delta Q_s = -\mu_L/3$, The estimate obtained from this regression, $\hat{\lambda}_{imp}$ will converge to

$$\begin{aligned}
\lambda_{imp} &= \frac{Cov\left(\Delta\bar{\mu}_s, \Delta Q_s\right)}{Var\left(\Delta Q_s\right)} = \frac{Cov\left(\frac{\mu_{Zs}-\mu_{Ls}}{3}, -\frac{\mu_{Ls}}{3}\right)}{Var\left(-\frac{\mu_{Ls}}{3}\right)} \\
&= \frac{Var\left(\mu_{Ls}\right) - Cov\left(\mu_{Zs}, \mu_{Ls}\right)}{Var\left(\mu_{Ls}\right)} \\
&= 1 - \rho,
\end{aligned}$$

where we have assumed that $Var\left(\mu_L\right) = Var\left(\mu_Z\right)$ in the final step. This expression shows that when $\rho > 0$, as is true empirically for teachers with non-missing VA, imputing 0's will bias λ_{imp} downward relative to 1, i.e. overstate the degree of forecast bias $B = 1 - \lambda$.[12]

Rothstein (2016, Section 3.2) recognizes that this imputation procedure generates bias, but argues that the bias from assigning all teachers a VA of 0 is small. To assess the validity of this claim, we construct a simple Monte Carlo simulation in which VA is missing at random for 20% of teachers and has a correlation of $\rho = 0.2$ across teachers within schools, matching the empirical values in the New York data (see Appendix C for the code). When missing data are excluded,

[12]The fact that $\rho > 0$ implies that one may be able to obtain a better forecast of teacher VA by using information about their peers' performance. This is not inconsistent with CFR's conclusions that standard VA estimates, which are based on a teacher's own performance, are forecast unbiased. As CFR-I (Section I.C) emphasize, their analysis does not show that standard VA estimates are *optimal* forecasts of teacher effectiveness given all available information; it only establishes that these VA estimates are unbiased forecasts.

a regression of changes in mean test scores on changes in mean VA in this simulation yields an estimate of λ that is not significantly different from 1, as expected. In contrast, imputing VA of 0 for teachers with missing VA yields an estimate of $\lambda_{imp} \simeq 0.93$. Hence, imputing teacher VA leads to a downward-biased estimate of λ by about 7%, similar to the magnitude by which the empirical estimates of λ change when imputed data are included in both the New York and North Carolina data.

The analysis above gives one simple example to illustrate why Rothstein's preferred imputation procedure will generally yield inconsistent estimates of forecast bias. This is just one of many potential ways in which the parametric modeling assumptions necessary to impute VA might be violated. Another violation, discussed by BKS (2016), arises from the fact that the teachers with missing data – who tend to be less experienced – have lower true VA on average (based on estimates that do not leave two years of data out). This again leads to bias when one imputes a value of 0. The broader point is that any imputation procedure – imputing 0 for all teachers or more sophisticated procedures – will generate measurement error in teacher VA unless the imputations perfectly mirror the data generating process for teacher VA, thereby producing biased estimates of λ.

I.C Discussion

There are two approaches to dealing with the selection bias that could arise from missing estimates of teacher value-added. One approach is to impute VA based on a specific model of the VA distribution for teachers with missing data. Although theoretically appealing, this approach is prone to errors because one must make strong, untestable parametric assumptions about the distribution of VA for teachers with missing data. An alternative, non-parametric approach is to restrict the sample to school-grade-subject-year cells without missing data and assess how the estimates change. This approach consistently yields estimates of forecast bias close to zero in the New York, North Carolina, and Los Angeles datasets. In our view, this evidence shows conclusively that missing data does not affect estimates of forecast bias significantly.

Since imputing VA of 0 to teachers with missing data yields inconsistent estimates of forecast bias, we do not consider specifications with imputed VA data further in this paper.[13]

[13]Rothstein suggests that there is an interaction between his concern regarding missing data and his concern about the correlation of teacher VA with prior test scores (which we turn to below). In particular, he argues that the prior score correlation is "mostly an artifact of CFR-I's sample construction, which excludes a non-random subset of classrooms [due to missing VA data]." However, this hypothesis is contradicted by his own estimates showing that controlling for prior scores reduces the estimate of λ by a similar magnitude even in the subsample of schools that

II Using Prior Test Scores for Placebo Tests

The second issue raised in Rothstein (2016) concerns the identification assumptions underlying CFR's quasi-experimental switchers design. Both Rothstein (2016) and BKS (2016) show that regressing changes in mean scores in the *prior* grade on changes in mean teacher VA across cohorts yields a positive coefficient, a result we confirm in our own data. Rothstein argues that this "placebo test" shows that CFR's quasi-experimental teacher switching design is invalid, as changes in current teacher quality cannot have a causal effect on students' past test scores. He then shows that if one controls for the differences in prior test scores across cohorts, VA estimates exhibit greater forecast bias than suggested by CFR's analysis. In particular, controlling for lagged test scores in the North Carolina yields an estimate of $\lambda = 0.933$ (Rothstein 2016, Table 3, Column 2). This estimate differs from CFR-I's (Table 4, Column 2) estimate – which does not control for lagged scores – of $\lambda = 0.957$ by 0.024 (2.4% forecast bias), a difference that is less than the standard error of CFR's estimate.

In this section, we explain why the correlation between changes in prior (lagged) scores and current teacher VA is consistent with the assumptions underlying CFR's research design and in fact should be expected because VA is estimated using prior test scores. We first demonstrate using stylized examples and Monte Carlo simulations that changes in current teacher VA will generally be correlated with changes in prior scores even when the quasi-experimental research design yields valid estimates of causal effects.[14] We then return to the New York data and show that excluding variation from teachers who switch across grades within a school eliminates the correlation with prior scores but does not affect our baseline estimates of forecast bias. These empirical results confirm that the relationship between changes in VA and lagged scores is spurious and not driven by selection bias. Finally, we show that Rothstein's approach of controlling for lagged test scores yields inconsistent estimates of forecast bias, unlike CFR's original implementation of the quasi-experimental design.

We begin with a simple example that illustrates the fundamental problem in using prior test scores to conduct placebo tests in value-added models. We then present the simulation results and turn to the empirical results.

have no missing teacher VA data (Rothstein 2014, Appendix Table 5, Column 4, Panels B and C). We therefore address the missing data and prior score issues independently.

[14]This discussion parallels an earlier debate about the test for selection bias in VA estimates using prior scores proposed by Rothstein (2010). Goldhaber and Chaplin (2015) and Chetty, Friedman, and Rockoff (2016) present theoretical and simulation-based evidence that this test can falsely detect bias in VA estimates even under random assignment of students to classrooms for reasons related to those discussed below.

II.A Stylized Example

In Table 2a, we present an example that illustrates why fluctuations in test scores due to shocks can produce a spurious correlation between changes in teacher VA across cohorts and changes in prior test scores. Consider two adjacent cohorts of 5th grade students in a given school and focus on their math test scores. Cohort 1 attends 5th grade in 1994 and cohort 2 in 1995. A 5th grade math teacher who has worked for several years switches to a different grade in the school after the 1994 school year, and is replaced with a new math teacher in 1995. Assume that the students in the two cohorts do not differ in their latent ability on average, so CFR's approach of identifying the effects of teacher VA based on differences in 5th grade test scores across cohorts yields unbiased estimates.

Kane and Staiger (2001) document that there are significant fluctuations in test scores across years within schools because of shocks that are correlated across students (e.g., because the math curriculum in a school happens to be well-aligned with the standardized math test administered in a given year). To understand the impacts of such shocks in our example, suppose the school had an idiosyncratic non-persistent positive shock to math scores in 1993, but that otherwise scores and teacher quality remained stable over time at some level that we normalize to zero. This positive shock causes cohort 1 (who is in 4th grade in 1993) to have higher 4th grade math scores than cohort 2 (who is in 4th grade in 1994). Thus, from the perspective of 5th grade, the change in *prior* test scores across the 1994-1995 cohorts is negative because of the positive shock in 1993.

Now consider the implications of the shock in 1993 for teachers' VA estimates. Since the departing 5th grade teacher was teaching in the school in 1993 when the positive shock occurred, her estimated VA will be higher because her VA is estimated using test score data from 1993 (and other prior years). Hence, the positive shock in 1993 will increase the estimated VA of the departing teacher relative to the new teacher, whose VA was *not* estimated using data from 1993 because she was not in the school at that time. Thus, the positive shock in 1993 also induces a negative change in mean teacher VA between the 1994-1995 cohorts.

Putting together the changes in test scores and changes in VA, we see that random school-level shocks in 1993 will induce a *positive* correlation between changes in mean teacher VA in 5th grade between 1994-1995 and prior test scores for 5th graders. Thus the "placebo test" using prior scores rejects the quasi-experimental design even though it is actually valid. The fundamental problem is that the prior test score data (in this example, from 1993) is used both to evaluate the validity of

the design and to estimate VA, so transitory shocks enter both the VA estimates and prior scores.

If the shocks to test scores were independent across years, as assumed in the example in Table 2a, one could eliminate the spurious correlation with prior scores by leaving out additional years of data when estimating VA. For example, consider a leave-three-year-out estimate of VA – i.e., an estimate of VA that excludes data from 1993-1995. Such a leave-three-year-out VA estimate would be unaffected by shocks in 1993 and thus would not be correlated with prior test scores in the simple example. However, the more plausible and empirically relevant scenario is one in which shocks are serially correlated across years, in which case a three-year-leave-out does not eliminate the spurious correlation with prior scores. We illustrate this point in Table 2b, which presents an example of a serially correlated shock. In this example, there is positive shock to mean test scores of +1 in 1992. This shock dissipates to +0.5 in 1993 and disappears entirely by 1994. Here, mean prior (4th grade) test scores are 0.5 units higher for the 5th graders in 1994 relative to 5th graders in 1995. And the shock of +1 in 1992 increases the leave-three-year-out VA estimate of the departing 5th grade teacher, but has no effect on the VA of the entering teacher. As a result, leaving out additional years of data when estimating VA, as suggested by Rothstein (2016, Online Appendix), does not eliminate the spurious correlation between changes in VA and changes in prior test scores.[15] Unless one knows the precise correlation structure of shocks to test scores, one cannot construct a VA estimate that is guaranteed to be orthogonal to noise in test scores in any period.

Note that in both of these examples, the change in 5th grade math scores across cohorts is unaffected by the prior shocks, so the shocks do not affect the relationship between the change in VA and the change in *current* test scores. Hence, the CFR design yields unbiased estimates of λ in both of these examples despite failing the prior score "placebo test." If the shocks were more persistent and decayed differentially across later cohorts, they could induce violations of the identification assumption required for CFR's design, which requires that student unobservables are balanced across the two cohorts (Assumption 3 in CFR-I). We view such differential shocks as one of many unobservables that could potentially vary across cohorts and violate the identification assumption underlying the quasi-experimental design. This observation underscores an important point: the examples above simply demonstrate that Rothstein's analysis of prior test scores is not informative about the validity of the quasi-experimental design. Of course, this fact does not automatically imply that the CFR design is actually valid. To assess the validity of the design, one

[15] In this stylized example, in which shocks dissipate within three years, a leave-four-year-out estimate of VA would eliminate the spurious correlation. In general, however, leaving out additional years of historical data will not solve the problem because the shocks may persist over longer horizons.

needs to use other approaches, such as the tests implemented in CFR-I, which we review briefly at the end of this section.

II.B Simulation Results

To verify the logic of the simple example above, we present simulation results which illustrate that one will find correlations with prior test scores even when the quasi-experimental design yields consistent estimates of the degree of forecast bias. The code for this simulation is provided in Appendix C.

Our simulation is based on a simplified version of the statistical model in CFR-I. We consider only one subject and ignore drift in teacher quality across years. We also ignore heterogeneity across students in terms of ability, so that variation in test scores is driven solely by teacher value-added and shocks to test scores. We allow for three types of shocks to test scores, each of which are present in the data: idiosyncratic student-level shocks, class-level shocks, and a serially correlated shock at the school-year level. Teachers switch among school-grade cells, generating the variation that drives our quasi-experimental design. Teachers are randomly assigned to students, so the true causal effect of VA on test scores is $\lambda = 1$.

We use the simulated data to estimate forecast bias and correlations with prior test scores using the teacher switchers design. The results of this analysis are presented in Table 3. The first row of this table reports estimates from OLS regressions of changes in mean current test scores across cohorts (ΔA_t) on changes in mean VA across cohorts (ΔVA_t). The second row reports estimates from OLS regressions of changes in prior test scores (ΔA_{t-1}) on ΔVA_t. Each column reports estimates from a different specification.

In Column 1, we estimate VA using a two-year-leave-out, as in CFR-I. That is, we exclude data from years t and $t-1$ when estimating VA and calculating the change in mean VA between school years t and $t-1$ in each school-grade cell. We then regress the change in test scores (ΔA_t) on the change in value-added to obtain our baseline quasi-experimental estimate. This regression yields an estimate of $\lambda \simeq 1$ (Panel A, Column 1), confirming that the quasi-experimental design recovers the true value of λ in this simulation. Panel B of Column 1 shows that regressing the change in lagged scores (ΔA_{t-1}) on ΔVA_t yields a coefficient of 0.138. That is, changes in VA are spuriously correlated with changes in prior test scores even though teachers are randomly assigned to students.

In Column 2, we follow Rothstein's (2016, Online Appendix) proposal of leaving out additional years of data when estimating VA. In particular, we use a three-year-leave-out by excluding years

$t-2$, $t-1$, and t when estimating the change in mean VA between years t and $t-1$. The coefficient on the prior score (Panel B) remains positive at 0.097 in this specification, showing that leaving out additional years of data does not eliminate the spurious correlation with prior scores. Meanwhile, the coefficient on current scores remains at 1 in this specification.

In our simulation, the only source of the correlation between changes in prior scores and current VA is serially correlated school-year shocks (of course, more generally there may be many other shocks that produce such a correlation). We can eliminate these shocks by adding school-year fixed effects to our regression specifications.[16] Column 3 shows that including school-year fixed effects leaves the coefficient on current scores essentially unchanged ($\lambda = 0.972$ with a standard error of 0.018), but eliminates the "effect" of teacher VA on prior test scores, where we now obtain a statistically insignificant coefficient of 0.009. This confirms that the relationship between prior scores and current VA is in fact driven by school-year shocks, with no consequence for the estimated effect on current scores.

These simulations demonstrate that lagged test scores are uninformative about the degree of bias in the quasi-experimental design. Even in a setting where teachers are randomly assigned to students and the CFR-I design recovers the correct coefficient on the effect of value-added on current scores, changes in measured VA and changes in lagged scores are positively correlated. Moreover, the correlation between prior test scores and changes in mean VA fluctuates substantially across specifications that use different sources of variation even though the estimate of the parameter of interest (λ) remains stable.

II.C Empirical Results

The simulations above establish that the correlation between changes in prior test scores and mean teacher VA provide no definitive evidence for or against the validity of the quasi-experimental design. Nevertheless, one may still be concerned that the correlation between prior scores and changes in VA in the data is driven by selection bias rather than spurious effects. To address such concerns, we return to the New York data and show that accounting for some simple mechanical effects eliminates the correlation between changes in lagged scores and current teacher VA but does not affect the original estimate of forecast bias.

Table 4 presents a set of variants of the baseline specification used in Table 4 of CFR-I to estimate forecast bias. As in Table 3, Panel A reports estimates from regressions of changes in

[16] Because we only consider one subject in our simulation, school-year fixed effects are equivalent to the school-year-subject fixed effects that we use in our empirical analysis below.

mean current test scores across cohorts (ΔA_t) on changes in mean VA across cohorts (ΔVA_t), while Panel B reports estimates from regressions of changes in mean prior test scores (ΔA_{t-1}) on ΔVA_t. In addition to the coefficient and standard error, we also report p-values from tests of equality with our baseline estimate ($\lambda = 0.957$) for the regressions with current scores in Panel A and tests for equality with 0 for the regressions with lagged scores in Panel B.

We begin in Column 1 by replicating the baseline specification used by Rothstein (2016, Table 2, Column 1), which includes school-year fixed effects. For current scores, we obtain a coefficient on changes in teacher VA of $\lambda = 0.957$, replicating the estimate reported in CFR-I (Table 4, Column 2). For prior scores, the corresponding regression coefficient is 0.171, an estimate that is significantly different than zero and similar in magnitude to Rothstein's estimate of 0.144 (Rothstein 2016, Table 2, Column 1). To assess whether this coefficient of 0.171 on prior scores reflects a spurious correlation as suggested by our simulations or selection bias as suggested by Rothstein, we implement three sets of specifications that isolate different sources of variation in mean teacher VA.[17]

1) Dropping Teachers who Follow Students. One direct way in which a teacher's VA can be correlated with students' prior test scores is if the teacher taught the same students in a previous year. Such "followers" create a correlation between changes in teacher VA and lagged scores across cohorts because noise in students' lagged scores directly enters these teachers' VA estimates and because these teachers have direct treatment effects on prior scores.[18] We remove the variation arising from followers by calculating mean VA for teachers in grade g in year t excluding teachers who taught in grade $g-1$ in the same school in year $t-1$, which we denote by VA_t^{nf}. We then replicate the specifications in Column 1 of Table 4, instrumenting for the actual change in VA (ΔVA_t) using ΔVA_t^{nf}.[19] The identification assumption required to obtain consistent estimates of λ using this IV approach is that changes in ΔVA_t^{nf} are orthogonal to changes in student unobservables, a variant of Assumption 3 in CFR-I.

Column 2 of Table 4 reports the resulting 2SLS estimates. The 2SLS regression coefficient for prior test scores falls to 0.052 in this specification, and is no longer statistically significant. Meanwhile, the 2SLS coefficient for current scores is 0.923, and we cannot reject equality with the

[17]The purpose of these alternative specifications is not to change CFR-I's original design and search for a suitable specification; rather, these alternative specifications provide insight into whether the prior score correlation biases the estimates obtained from the baseline specifications in CFR-I.

[18]See Appendix A for further details on the precise mechanics of this effect.

[19]The first-stage coefficient in this 2SLS regression is 0.97, as expected given that very few teachers follow their students across grades.

baseline estimate of 0.957. In the North Carolina data, Rothstein (2016, Appendix Table A8, Row 3) finds that the same no-followers specification reduces the coefficient for lagged scores by about 50% as well, from 0.144 to 0.08, while the coefficient for current scores remains relatively stable at $\lambda = 1.00$. These results show that simply excluding teachers who follow students when computing mean VA eliminates most of the correlation with prior scores without changing the baseline estimate of forecast bias appreciably.

2) Distinguishing Within-School and Between-School Switchers. Building upon the approach of excluding teachers who follow students, in Column 3 of Table 4, we exclude all teachers who switch across grades within the school when constructing mean VA in each cohort. Again, we instrument for the change in actual mean VA with this modified change in mean VA excluding within-school switchers. The 2SLS coefficient on the change in teacher VA in a regression of current scores is 0.968, not significantly different from our baseline. The 2SLS coefficient for lagged scores is 0.031 and not significantly different from zero. This suggests that teachers who switch within the school face shocks that also affect their students' prior performance (e.g., school-specific shocks). We explore this point further in Column 4, where we define mean VA purely using teachers who switch within the school – that is, the complement of the set of teachers used to define mean VA in Column 3. Column 4 shows that instrumenting for the actual change in VA using only the variation driven by *within-school* movement (including followers across adjacent grades) yields a much larger coefficient on prior scores of 0.258. Yet the coefficient on current scores remains stable at 0.950, implying that the correlation with lagged score has no bearing on the estimate of λ. Together, these specifications suggest that, at least in the New York data, much of the correlation between prior scores and teacher VA is generated by correlated shocks at the school level that ultimately have little or no impact on the estimates of forecast bias.

3) Accounting for School-Year-Subject Shocks. In Column 5 of Table 4, we include school-subject-year fixed effects (rather than school-year fixed effects) and replicate the specification in Column 2. This specification nets out school-year-subject shocks, which are highly statistically significant in our data ($p < 0.001$). In this specification, we obtain an estimate of $\lambda = 0.942$ for current test scores and a coefficient of $-.023$ for lagged scores. Rothstein (2016, Appendix Table A8, Row 5) also estimates this specification and obtains a coefficient for current scores of 1.03, *identical* to his baseline estimate, and a statistically insignificant coefficient for lagged scores of 0.05.

BKS (2016, Table 4) report analogous results in the Los Angeles data. For example, removing

within-school switchers reduces the coefficient on prior scores in their data to an insignificant 0.049, while the coefficient on current scores remains at 0.963. The similarity of the results across all three datasets suggests that these patterns are a robust feature of the data generating process across school districts rather than a statistical anomaly in any one sample.

The central lesson of Table 4 is that the coefficient on current scores (λ) is very stable across specifications that isolate different pieces of variation in teacher VA, while the coefficient on prior test scores fluctuates substantially. If the relationship with lagged scores in the baseline specification did in fact reflect true differences in student ability across cohorts, one would expect the estimates of λ to fall substantially when one isolates a source of variation in VA that is uncorrelated with prior test scores. Therefore, these results imply that the correlation with lagged scores in the data reflects spurious noise and does not bias CFR-I's baseline estimates of λ.

II.D Controlling for Prior Test Scores

Rothstein (2016, Section 3) argues that it is preferable to control for changes in lagged scores ΔA_{t-1} when regressing ΔA_t on ΔVA_t to estimate λ. Although controlling for pre-determined prior scores seems like it should improve estimates, in this subsection we show that controlling for ΔA_{t-1} will typically *increase* bias in the estimate of λ in this setting, for two reasons.

First, because much of the correlation between ΔA_{t-1} and ΔVA_t is driven by teachers who follow students across grades, lagged test scores are endogenous to changes in teacher VA. Including such an endogenous control naturally generates bias because part of the causal effect of ΔVA_t is picked up by ΔA_{t-1}. The pitfalls of including endogenous controls – termed "bad controls" by Angrist and Pischke (2009) – are well known and so we do not discuss them further here.

Second, controlling for ΔA_{t-1} creates bias when the correlation between ΔA_{t-1} and ΔVA_t is largely driven by transitory shocks rather than true differences in student ability, as appears to be the case empirically. To see this, let ψ denote the coefficient from a regression of ΔA_t on ΔA_{t-1}; we estimate $\psi = 0.64$ in our data, while Rothstein estimates $\psi = 0.675$. These large coefficients arise because the correlation between ΔA_t and ΔA_{t-1} is driven mostly by variation in student ability across cohorts, which is highly persistent, rather than transitory fluctuations in test scores.

Suppose the variation in ΔA_{t-1} that is correlated with ΔVA_t is driven by transitory shocks to prior test scores that have no impact on ΔA_t, as in the examples in Table 2a and 2b. In this case, the value of ψ that applies to the variation in ΔA_{t-1} that matters – i.e., the portion that is correlated with ΔVA_t – is $\psi_C = 0$. Here, we know that the baseline OLS regression that

does *not* control for ΔA_{t-1} yields a consistent estimate of λ. Therefore, the omitted variable bias formula implies that controlling for ΔA_{t-1} will yield an estimate of λ that is *downward-biased* by$\Delta\lambda$ =-$\psi\theta$, where θ denotes the coefficient from a regression of ΔA_{t-1} on ΔVA_t. In our baseline specification, we estimate $\theta = 0.171$ (Column 1 of Table 4, Panel B), while Rothstein estimates $\theta = 0.144$. Controlling for ΔA_{t-1} when $\psi_c = 0$ would therefore produce a downward bias of $\Delta\lambda$ =-$\psi\theta \simeq -0.144 \times 0.675 = -0.0973$, matching the change in λ of -0.097 that Rothstein finds when controlling for lagged scores (Rothstein 2016, Table 3a, Columns 1 and 2). Intuitively, by controlling for ΔA_{t-1}, one falsely attributes part of the difference in current test scores ΔA_t to differences in student ability when those differences are actually due to transitory shocks that enter both ΔVA_t and ΔA_{t-1} but have no bearing on ΔA_t.

More generally, including ΔA_{t-1} as a control yields biased estimates of λ if ψ_c, the persistence of the shocks that drive the correlation between ΔA_{t-1} and ΔVA_t, differs from $\psi \simeq 0.66$. The examples and simulations above demonstrate that ψ_C differs from ψ as soon as one incorporates empirically relevant features such as transitory school-year-subject shocks.

Of course, the fact that ψ_C differs from ψ does not necessarily mean that *excluding* ΔA_{t-1} from the control vector (which is equivalent to assuming $\psi_C = 0$) will yield unbiased estimates. Since ψ_C is unknown, the most definitive, non-parametric approach to assessing whether the correlation between VA and lagged scores generates bias is to isolate sources of variation in teacher VA that *eliminate* the correlation between ΔA_{t-1} and ΔVA_t and evaluate whether the estimates of λ change. This is precisely what we did in Table 4 above. The fact that the estimates of λ remain unchanged in subsamples where ΔVA_t is uncorrelated with ΔA_{t-1} directly shows that the correlation between ΔVA_t and ΔA_{t-1} in the baseline specification has no bearing on the estimate of λ.

Building on this logic, the evidence in Columns 2-5 of Table 4 can be used to infer that ψ_C is much closer to 0 than 0.66. In Panel C of Table 4, we calculate the predicted value of λ that one would obtain starting from the baseline estimates of $\lambda = 0.957$ and $\theta = 0.171$ in Column 1 of Table 4 under the dubious assumption that $\psi_C = \psi = 0.66$. We compute these predictions as $\lambda_p = 0.957 - 0.66 \times (0.171 - \theta)$, where θ is the coefficient from the regression of ΔA_{t-1} on ΔVA_t reported in Panel B in each column. Because the relationship between ΔVA_t and ΔA_{t-1} changes substantially across specifications, the predicted values assuming $\psi_C = \psi = 0.66$ vary from $\lambda_p = 1.015$ in Column 4 (where we use variation from within-school switchers) to $\lambda_p = 0.829$ in Column 5 (where we exclude followers and use school-year-subject fixed effects). Yet the actual estimates of λ in these specifications, shown in Panel A, hardly change ($\lambda = 0.950$ vs. $\lambda = 0.942$) –

as one would expect if $\psi_C = 0$. These empirical results are therefore much more consistent with the view that $\psi_C = 0$ than $\psi_C = 0.66$, i.e. that the relevant variation in ΔA_{t-1} is driven by transitory shocks rather differences in student ability. It follows that *excluding* ΔA_{t-1} as a control – precisely as in CFR's baseline design – yields unbiased estimates whereas including it does not.

In sum, our application is very different from Rothstein's (2016, p12) analogy to a randomized experiment where one detects imbalance in predetermined covariates and attempts to correct for this imbalance by controlling for the covariates ex-post. In our setting, we fully expect imbalance in ΔA_{t-1} across school-grade-year cells with different values of ΔVA_t because of the way in which VA is estimated, as illustrated by our simulations. Because this imbalance is due to noise rather than persistent differences in student ability and because prior scores are partly endogenous to changes in ΔVA_t, controlling for ΔA_{t-1} yields biased estimates.

II.E Discussion

The analysis in this section has three important implications.

First, regressions of changes in prior test scores on changes in mean teacher VA can detect "placebo effects" even when the underlying research design is valid. Because teacher VA is estimated using data from the same environment that students were previously in, serially correlated shocks to test scores will enter both VA estimates and students' performance in previous years.

Second, correcting for some obvious sources of such shocks – for instance, by dropping teachers who follow students or excluding within-school switchers when estimating mean teacher VA – eliminates the correlation between prior test scores and changes in current teacher VA but has no significant effect on the estimates of forecast bias in the New York, North Carolina, and Los Angeles datasets. These results confirm that the prior score correlation has no bearing on empirical estimates of forecast bias in practice. Note that these results should not be interpreted as implying that the lagged score test becomes valid after implementing "corrections" such as dropping within-school followers.[20] The point is not that the quasi-experimental specification needs to be modified to eliminate the correlation with prior test scores; it is that the prior test score "placebo test" is invalid and should not be used to assess the research design.

Third, controlling for lagged test scores yields biased estimates of the effects of teacher VA because lagged scores are partly endogenous to current teacher VA (when some teachers follow students) and because the variation in lagged scores appears to be primarily driven by noise rather

[20]For example, there could be correlated shocks across nearby schools, which would continue to create a spurious correlation between teacher VA and prior scores even after removing within-school switchers.

than persistent differences in student ability.

The fundamental source of all of these issues is that, unlike in a standard treatment effects setting where the treatments are exogenously observed, the "treatment" in VA models is itself estimated from the data. As a result, standard intuitions that prior values can be used to implement placebo tests or that one can obtain consistent estimates by controlling for prior values fail.[21]

Of course, the fact that prior test scores do not provide any information about the validity of CFR's quasi-experimental design does not mean that the design is valid. Instead, it means that one must use alternative tests to assess the validity of the design. In CFR-I, we implemented a series of such diagnostic tests. For example, we showed that changes in teacher VA across cohorts are uncorrelated with cross-cohort changes in parental characteristics such as income (see CFR-I Table 3). We also analyzed the effects of changes in teacher VA in a given subject (e.g., math) on test scores in the *other* subject (e.g., English). We found that cross-cohort changes in mean teacher VA in one subject are unrelated to changes in contemporaneous test scores in the other subject when students have different teachers in the two subjects (CFR-I Table 4, Column 5). Moreover, changes in prior test scores in the other subject are also uncorrelated with changes in mean teacher VA in a given subject.

Parental characteristics and performance in other subjects are both very highly correlated with students' test scores, and hence are likely to pick up any latent differences in student ability across cohorts. The fact that differences in teacher VA are uncorrelated with parental characteristics and other-subject achievement therefore provides further evidence that the lagged score correlation in own-subject achievement arises from spurious shocks that enter VA estimates rather than differences in latent student ability. Rothstein does not present any arguments as to why the alternative placebo tests implemented by CFR-I fail to detect selection bias and are less credible than analyzing prior scores. More generally, as we noted in CFR-I, the diagnostic tests show that any violation of the teacher switchers design "would have to be driven by unobserved determinants of test scores that (1) are uncorrelated with parent characteristics, (2) are unrelated to prior test scores and contemporaneous test scores in the other subject, and (3) change differentially across grades within schools at an annual frequency. We believe that such sorting on unobservables is implausible given the information available to teachers and students and the constraints they face in switching across

[21] This problem is not unique to the research design used by CFR; it applies to value-added models more generally. For example, school-year shocks will induce a correlation between VA estimates and prior test scores in a cross-section of classrooms – as documented e.g. in Rothstein (2010) – even if student unobservables are balanced across classrooms (Chetty, Friedman, Rockoff 2016).

schools at high frequencies."

III Estimating Teachers' Long-Term Impacts

The third issue raised in Rothstein (2016) concerns CFR-II's estimates of teachers' impacts on students' earnings. CFR-II report estimates using both OLS regressions and the quasi-experimental teacher switching design. Rothstein obtains similar estimates in the North Carolina data when examining comparable outcomes. However, he argues that both the OLS and quasi-experimental estimates are upward biased, because of two separate issues.

First, Rothstein points out that CFR-II's OLS regression estimates, like any OLS regression estimates, can only be interpreted as causal effects conditional on the assumption of selection on observables (i.e., that the covariates in the regression fully account for any confounding factors). We fully agree with this observation and emphasized on page 2 of CFR-II this is "a very strong assumption." Rothstein goes on to argue that CFR's method of controlling for observables – which uses within-teacher variation to identify the impacts of covariates – relies on particularly strong assumptions. He proposes an alternative approach that uses between-teacher variation for identification. In this section, we demonstrate using a Monte Carlo simulation that Rothstein's estimator yields biased results in a case where CFR-II's approach yields consistent estimates. Moreover, while our estimator is consistent under the assumptions laid out in CFR-II, we show that Rothstein's approach is internally inconsistent with the assumptions under which value-added is estimated.

Although Rothstein offers no definitive evidence that the identification assumptions made in CFR-II are violated, he is certainly correct to raise the possibility that the controls in our OLS regressions might not adequately account for omitted variable bias. Recognizing this concern, we reported estimates that do not rely on the assumption of selection on observables, instead using our quasi-experimental teacher switching research design (CFR II, Section IV). We found that this method yields very similar estimates of teachers' long-term impacts to OLS for the outcomes for which we have adequate precision (e.g., college attendance).

Rothstein finds similar results in data from North Carolina, but, returning to the issue of correlations with prior test scores discussed in Section 2 above, he argues that the quasi-experimental estimates of long-term impacts might be upward biased. He shows that one obtains smaller estimates of long-term impacts when controlling for prior scores. For the same reasons as those discussed in the previous section, prior scores are improper controls that generate bias by attenuating the effect of VA, and the switchers research design is valid as originally implemented in

CFR-II. Moreover, even after controlling for prior scores, the quasi-experimental estimates in the North Carolina data are statistically indistinguishable from estimates for comparable outcomes in the New York data.

Based on this analysis, we conclude that both the New York and North Carolina data support the hypothesis that teacher VA predicts teachers' long-term impacts. At a minimum, there is no evidence that the true impacts fall outside the confidence intervals reported in CFR-II.

We begin by summarizing Rothstein's empirical evidence on long-term impacts in the North Carolina data. We then address the issue of controlling for observables in OLS regressions and turn to the quasi-experimental estimates.

III.A Evidence from North Carolina Data

Rothstein's analysis of long-term impacts in North Carolina is not a direct replication of CFR-II because it examines different outcomes. CFR-II examine teachers' effects on college attendance rates and earnings, while Rothstein examines teachers' impacts on a set of outcomes measured in high school: graduation rates, plans to attend college, grade point average (GPA), and class rank. We focus on the high school graduation and planned college attendance outcomes here because they are most comparable to the outcomes in adulthood examined by CFR-II. Conceptually, teachers' effects on high school GPA and class rank could differ from their effects on the other outcomes for two reasons. First, a high VA teacher in elementary school might affect the high school a student attends or the classes a student takes – for instance, by inducing a student to take more advanced courses – which could lead to ambiguous effects on measured GPA and class rank. Second, prior work has shown that the impacts of educational interventions "fade out" when examining intermediate measures of academic achievement, only to re-emerge in adulthood (Deming 2009, Chetty et al. 2011, CFR-II). Because of these reasons, it is not clear that teachers' impacts on high school grades or ranks should parallel their impacts on outcomes in adulthood.[22]

When replicating CFR-II's OLS specification in the North Carolina data, Rothstein (2016, Table 5, Column 2) finds that a 1 SD increase in teacher VA increases high school graduation rates by 0.34 percentage points (s.e. = 0.04) and the fraction of children planning to attend college by 0.60 pp (s.e. = 0.07). When using CFR-II's quasi-experimental design, Rothstein (2016, Table 6, Column 2) obtains estimates of 0.38 pp (s.e. = 0.17) for high school graduation and 0.61 pp (s.e. = 0.24) for planning to attend college. These findings are closely aligned with CFR-II's results.

[22]Moreover, the GPA and class rank outcomes are missing for a majority of the students in the North Carolina data, so the quasi-experimental estimates for these outcomes suffer from an even greater lack of precision.

As in CFR-II, Rothstein's OLS and quasi-experimental estimates are very similar to each other. Moreover, Rothstein's estimate of the impact on the fraction of children who plan to attend college is consistent with CFR-II's estimate of 0.82 pp for actual college attendance.

Rothstein goes on to argue that both the OLS and quasi-experimental designs – which rely on different sources of variation and identification assumptions – are biased, for different reasons. We consider each of these two issues in turn.

III.B Controlling for Observables in OLS Regressions

CFR Two-Step Estimation Methodology using Within-Teacher Variation. We begin by reviewing the two-step approach to controlling for observable student characteristics used in CFR-I to estimate teacher value-added and CFR-II to estimate teachers' long-term impacts. For simplicity, we focus on a model without drift in teacher VA over time.

To estimate VA in CFR-I, we first construct student test score residuals, removing the effects of observable covariates. For student i in year t with teacher j, we first estimate the relationship between test scores A^*_{it} and a vector of covariates X_{it} using a fixed-effects OLS regression

$$A^*_{it} = \alpha_j + \beta X_{it}, \tag{2}$$

where α_j is a teacher fixed effect. Because we include teacher fixed effects, β is identified from variation across students taught *by the same teacher*. We then calculate the residualized test score

$$A_{it} = A^*_{it} - \beta X_{it}.$$

We use these residualized test scores to estimate VA $\hat{\mu}_{jt}$ using test score data from years excluding t, and then define forecast bias based on the coefficient from a univariate regression of the form:

$$A_{it} = \alpha_t + \lambda \widehat{\mu}_{jt},$$

as in equation (10) of CFR-I.

In CFR II, we use a parallel two-step approach to estimate teachers' long-term impacts. Consider the following structural model for earnings Y^*_i, which is a simplified version of the model in CFR-II (in equations (2) and (4) of the main text and (17) in CFR-II Appendix A) that ignores tracking effects and drift in teacher VA:

$$Y^*_i = \alpha + \tau_j + \beta^Y X_{it} + \eta_{it} \tag{3}$$

where τ_j represents teacher j's earnings value-added in year t – i.e., teacher j's direct impact on earnings. A teachers' earnings VA τ_j may be correlated with her test-score VA μ_j, but the correlation may be imperfect.

Our goal is to estimate the effect of a 1 unit increase in teacher's test-score VA μ_j on earnings, which we denote by κ. To estimate κ, we first regress each long-term outcome such as earnings Y_i^*on a vector of control variables, again including teacher fixed effects to identify from *within-teacher* variation across students:

$$Y_i^* = \alpha_j + \beta^Y X_{it}.$$

We then regress the residualized outcome variable $Y_{it} = Y_i^* - \beta^Y X_{it}$ on teachers' estimated test-score VA based on data excluding year t ($\hat{\mu}_{jt}$):[23]

$$Y_{it} = \alpha + \kappa \hat{\mu}_{jt} + \eta_{it} \tag{4}$$

without further controls. Under the assumption that η_{it} is orthogonal to μ_j (Assumption 2 in CFR-II), we can interpret $\kappa = \frac{Cov(Y_{it},\hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})} = \frac{Cov(\tau_j,\hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})}$ as an estimate of the causal effect of a 1 SD increase in a teacher's test-score VA on her students' earnings Y_{it}. Note that unlike in a standard partial-regression, we do not residualize the right hand size variable μ_j with respect to the covariates X_{it} when regressing earnings residuals Y_{it} on VA in (4). Doing so would yield an inconsistent estimate of κ because we estimated β^Y from within-teacher variation (see Appendix B for a proof).

The model for earnings in CFR-II and in (3) implies that the estimate of β^Y that one obtains from within-teacher variation in X_{it} is equivalent to the estimate of β^Y one would obtain from between-teacher variation (holding fixed true teacher quality). This is because our model assumes that differences in X_{it} lead to the same change in earnings for students taught by the same teacher as those taught by different teachers. Rothstein questions whether this assumption holds in practice and seeks to develop alternative estimators that do not rely on this assumption. Rothstein (2014, 2016) proposes two approaches to estimate κ – a multivariable OLS regression and a 2SLS estimator – that yield similar estimates. We consider each approach in turn.[24]

[23]For scaling purposes, CFR-II report the effect of a 1 unit increase in normalized teacher VA, defining the independent variable as $\hat{m}_{jt} = \widehat{\mu}_{jt}/\sigma_\mu$ where σ_μ is the variance of test-score VA. Because this just changes the coefficient estimates by a scalar, we simplify notation by writing the regressions as a function of $\widehat{\mu}_{jt}$ in this note.

[24]Other researchers have also considered 2SLS estimators analogous to that proposed by Rothstein (2014). Because the objective of this published exchange is to clarify the methodological issues that arise when estimating VA models for other interested researchers, we retain our discussion of the 2SLS estimator in Rothstein (2014) even though it is not included in Rothstein (2016).

Multivariable Regression using both Within- and Between-Teacher Variation. Consider the following OLS regression of earnings on the test-score VA estimate $\hat{\mu}_{jt}$ and the covariate X_{it}:

$$Y_i^* = \alpha + \kappa_R \hat{\mu}_{jt} + \beta^Y X_{it} + \eta_{it} \tag{5}$$

Rothstein (2016, Table 5, Column 5) shows that this multivariable regression yields smaller estimates of κ_R than those produced by the two-step estimator in (4) in the North Carolina data. The estimated effects on high school graduation and plans to attend college fall to 0.24 pp (s.e. = 0.04) and 0.41 pp (s.e. = 0.06), respectively – about 30% smaller than the baseline estimates produced by the two-step estimator. He argues that CFR's approach therefore overstates the long-term impacts of teachers.

Although intuitive at first glance, this multivariable regression in (5) does not yield a consistent estimate of κ because it does not control for differences in teachers earnings VA τ_j that are correlated with X_{it} and hence misestimates β^Y. For example, suppose that students with higher lagged test scores X_{it} are assigned to teachers with higher earnings VA τ_j. Because test score VA μ_j is only one component of earnings VA, the multivariable regression over-attributes explanatory power to lagged test scores, yielding an upward biased estimate of β^Y and a downward-biased estimate of κ_R. Intuitively, part of the relationship between lagged test scores and earnings is due to the fact that students with high prior test scores get teachers who have more positive effects on earnings. Since (4) does not fully control for teachers' effects on earnings, the coefficient on lagged test scores is upward-biased. This reduces the residual variation left to be explained by teachers' test score VA $\hat{\mu}_{jt}$ and thus yields a downward-biased estimate of κ.[25] More generally, a multivariable regression will underestimate the effect of VA on long-term outcomes because it will over-attribute variation in students' outcomes to the X's rather than teachers. Our approach of estimating β^Y using *within-teacher* variation was designed to resolve precisely this problem, as it mechanically eliminates any conflation of the effects of teachers with the effects of the control variables.

To illustrate this problem, we report estimates based on Monte Carlo simulations in Table 5 (see Appendix C for the code underlying this simulation). We use a simple data-generating process for scores and earnings in which teachers' earnings VA τ_j is correlated with a control variable X. We assume that the true effect of a 1 unit increase in VA on earnings is κ = \$100 and the true effect of a 1 unit increase in X is \$10. Column 1 of Table 5 shows estimates from CFR's approach

[25] Stated differently, because test-score VA $\hat{\mu}_{jt}$ captures only one component of teachers' earnings VA τ_j, the multivariable regression in (5) is effectively biased by measurement error in earnings VA that leads us to misestimate β^Y and κ.

of constructing residuals Y_{it} using within-teacher variation and regressing these residuals on VA estimates. As expected, we obtain an estimate of $\hat{\kappa} \simeq 100$ using this approach. Column 2 shows that in contrast, the multivariable regression in (5) yields an estimate of $\hat{\kappa} = 75$, 25% lower than the true κ. This is because the estimate of β^Y is \$66, higher than its true value of \$10 because it picks up teachers' effects on earnings.

Two-Stage Least Squares Estimates. To account for measurement error in test-score VA, Rothstein (2014) proposes the following regression specification, which replaces test-score VA $\hat{\mu}_{jt}$ with test score residuals A_{it}, using two-stage-least squares:

$$Y_i^* = \alpha + \kappa_{IV} A_{it} + \beta^Y X_{it} + \eta_{it} \tag{6}$$

He instruments for a student's test score residual A_{it} with the test-score VA estimate $\hat{\mu}_{jt}$. Column 4 of Table 5 replicates this 2SLS estimator and shows that it too yields a downward-biased estimate of κ and an upward biased estimate of β^Y ($\hat{\kappa} = 80$, $\hat{\beta^Y} = 54$). Conceptually, this 2SLS estimator does not fix the problem that teachers' true earnings VA is correlated with the covariates and thus continues to misestimate β^Y. The 2SLS estimator is simply the reduced-form estimate in (5) divided by the first-stage coefficient from a regression of A_i on $\hat{\mu}_{jt}$ and X_{it}, which we denote by ϕ. In our simulation, $\phi = 0.94$, and hence $\kappa_{IV} = \kappa_R/0.94$ is similar to the OLS estimate and substantially biased relative to the truth. More generally, the 2SLS estimator will yield a consistent estimate of κ only if $\phi = \kappa_R/\kappa$, and there is no reason for this equality to hold. Intuitively, the 2SLS estimator accounts for the effect of measurement error in estimating test-score VA, but it does not account for the fact that teachers' *total earnings* VA is not properly measured and controlled for in the multivariable regression. Hence, part of the effect of teacher VA is still attributed to X_{it} in the 2SLS regression. These results demonstrate that Rothstein's (2014, p29) statement that the IV estimator is "consistent under more general conditions than the restrictive assumptions required for consistency of CFR's two-step estimator" is not correct. On the contrary, it is not consistent under even the assumptions required for CFR's estimator.

Internal Consistency. A further conceptual problem with the multivariable regression in (5), whether estimated using OLS or 2SLS, is that it is inconsistent with the way in which teachers' test score VA $\hat{\mu}_{jt}$ is estimated. As noted above, in CFR-I we residualize student test scores using within-teacher variation to estimate the vector of coefficients on the controls. This approach again is designed to account for the fact that the covariates vary across teachers, and hence using the pooled variation to construct test score residuals would lead us to overstate the effects of the covariates on

test scores.[26]

Rothstein's approach to estimating long-term effects introduces an asymmetry between the way long-term impacts are estimated and teacher VA is estimated. If there are indeed significant differences in the within-teacher and between-teacher relationships between X and student's outcomes, then one cannot use within-teacher variation to residualize student test scores and estimate teacher VA to begin with. Since Rothstein's VA estimates are based on the methodology of CFR-I, his approach to estimating teachers' long-term impacts is internally inconsistent with his VA estimates. That is, VA is estimated under the assumption that the structural relationship between covariates and student outcomes is the same within and between teachers, but long-term impacts are estimated under the assumption that the two relationships differ. Regardless of which assumption one believes is correct, one cannot make different assumptions in different parts of the analysis.[27]

In sum, our two-step method yields an unbiased estimate of the effect of teacher VA on long-term outcomes (κ) under the assumptions of the model we specify in CFR-I and II, whereas the multivariable regression approach does not. Nevertheless, we still fully agree with Rothstein on the bigger-picture point that our OLS regression estimates rely on strong assumptions. The assumption that within-teacher and between-teacher variation yield similar effects of covariates on outcomes might not hold in practice, just as the assumption that selection on observables given the covariates we use may not hold in practice. Conceptually, we are just assuming a particular form of selection on observables, and as we stressed in CFR-II, the OLS regression estimates can only be interpreted as causal effects if this assumption holds. The only way to evaluate the validity of this assumption is to use a research design that generates variation in VA orthogonal to observables. In the case of test scores, the fact that our quasi-experimental tests generate a prediction coefficient of $\lambda \simeq 1$ (no forecast bias) implies that our method of controlling for observables using within-teacher variation yields unbiased estimates of teachers' impacts of test scores. We can evaluate whether the same holds for outcomes such as college attendance by using the quasi-experimental design to estimate the long-term impacts of teacher VA, which we turn to next.

[26] For example, if minority students tend to have lower-VA teachers, then we would over-estimate the effects of race on test scores if we did not include teacher fixed effects when estimating the relationship between race and test scores using (2), because part of the coefficient on race would include the effects of teacher quality.

[27] Rothstein (2014, p28) suggests that one solution to this problem is to estimate VA conditional on covariates, which is analogous to using both within- and between-teacher variation to residualize *both* test scores and earnings. We pursued this approach in the working paper version of CFR-II (NBER wp 17699). In that paper, we obtained estimates of long-term effects of teacher VA that were roughly similar to those reported in the published version of CFR-II. However, as Rothstein noted in his original referee report on CFR-II, this approach to residualization using pooled variation does not yield consistent estimates of teacher effects, which is why we changed our estimation methodology in the revision.

III.C Quasi-Experimental Estimates

In Table 5 of CFR-II, we report estimates of teachers' effects on college attendance rates using our cross-cohort teacher switchers design. We find that these estimates of the long-term effects of VA are quite similar to the OLS estimates. The estimated impact of a 1 SD increase in teacher VA on college attendance rates is 0.82 pp using OLS (CFR-II, Table 2) compared with 0.86 pp using the quasi-experimental design. As noted above, Rothstein finds that the quasi-experimental estimates match the OLS estimates in the North Carolina data as well.

However, Rothstein argues that the cross-cohort estimates of teachers' long-term impacts are biased for the same reason as in Section II above: changes in teacher VA are correlated with prior test scores. Controlling for prior scores in the quasi-experimental design, the estimated effect on high school graduation falls from 0.38 pp to 0.26 pp (s.e. = 0.17) and the estimated effect on plans to attend college falls from 0.61 pp to 0.41 pp (s.e. = 0.24) (Rothstein 2016, Table 6, Columns 2 and 3). Based on these findings, Rothstein concludes that "no strong basis for conclusions about the long-run effects of high- vs. low-VA teachers, which in the most credible estimates are not distinguishable from zero."

We disagree with this conclusion for two reasons. First, following exactly the same reasoning as in Section II.D, the estimates that control for prior test scores are inconsistent because part of the effect of changes in VA is incorrectly attributed to differences in prior test scores, which appear to be driven by noise rather than persistent differences in students' ability. CFR-II present a set of alternative specification tests that do not suffer from the problems inherent in controlling for prior test scores. For example, they show that changes in leads and lags of teacher VA for adjacent cohorts do not predict changes in long-term outcomes (CFR-II, Figure 6) and that changes in predicted outcomes based on parental characteristics are uncorrelated with changes in teacher VA across cohorts (CFR-II, Table 5, Column 5). As in his critique of CFR-I, Rothstein does not explain why these alternative diagnostic tests implemented by CFR-II fail to detect selection bias and are less credible than analyzing prior scores.

Second, even if one takes the estimates with prior controls at face value, they are not statistically distinguishable from Rothstein's baseline estimates or from CFR-II's estimates for the college attendance outcome. Although the point estimates are about 30% smaller than the baseline estimates obtained without prior score controls, they differ from those estimates by less than one standard error. Thus, the fact that they "are not distinguishable from 0" as Rothstein emphasizes

is simply a statement about the lack of statistical power in Rothstein's sample rather than evidence that teachers' long-term effects are actually close to 0.

III.D Discussion

Stepping back from the details of the estimation procedure, our reading is that Rothstein's findings strongly support the view that teachers have substantial long-term impacts on high school graduation rates and plans to attend college, the outcomes most comparable to those studied by CFR-II. All of Rothstein's OLS specifications yield highly significant positive estimates for the impacts of teacher VA on these outcomes. Rothstein's baseline quasi-experimental specifications also yield significant positive estimates that match his OLS estimates, while his estimates that control for prior scores yield positive estimates that are not distinguishable from the baseline results.

In our view, the similarity of the quasi-experimental and OLS estimates in both the New York and North Carolina data constitutes particularly strong evidence that high-VA teachers create substantial long-term gains. Because the two estimates rely on entirely different sources of variation and identification assumptions, it would be surprising if they were both biased yet happened to yield such similar results.

Rothstein's preferred point estimates for these outcomes are about 30% smaller than the estimates he obtains when using CFR-II's specifications. But such differences in magnitudes fall within the confidence intervals for the quasi-experimental estimates of long-term impacts reported in CFR-II. Even if teachers' true long-term impacts are in fact 30% smaller than suggested by CFR-II's point estimates, it would not alter CFR's main qualitative conclusion that high-VA teachers improve their students' outcomes in adulthood substantially.

IV Conclusion

Rothstein (2016) provides an exceptionally thorough replication and re-examination of CFR-I and II that is a very useful contribution to the literature on teacher effectiveness. However, his three central critiques of CFR's methodology are not valid: his preferred method of imputing teacher VA generates bias, his proposed prior score placebo test falsely rejects valid research designs, and his alternative method of controlling for covariates yields inconsistent estimates of teachers' long-term effects. Excluding these incorrect results, his findings support – and in fact nearly duplicate – the results reported in CFR. Although we disagree with his conclusions, we are very grateful to Rothstein for helping us improve our own understanding of these issues through his insightful

comments and questions both in his formal referee reports and this followup response.

We conclude with two broader points about this debate. First, from an empirical perspective, Rothstein proposes parametric solutions for each potential problem he uncovers: imputation of missing data, adding a control for prior test scores, or changing the control vector in OLS regressions to estimate teachers' long-term impacts. These parametric approaches rely on strong assumptions that may not hold in practice, and in fact generate significant bias under plausible assumptions. We believe a better way forward, both in the current debate and in future work, is to obtain more definitive evidence using "non-parametric" methods that do not rely on such assumptions. For instance, in the cases of Rothstein's three critiques, we focus on samples with no missing data, isolate variation in teacher VA that generates no correlation with prior test scores, and use the quasi-experimental teacher switching design to estimate long-term effects. These non-parametric tests do not uncover any problems in CFR's methodology. Our view is that parametric approaches are undesirable unless one has an explanation for why the non-parametric evidence is flawed (e.g., why subsamples with no missing data exhibit no forecast bias).

Second, the broad message that we take from recent work is that forecast bias in standard value-added models is small, even if the exact magnitudes differ across studies. CFR-I estimate forecast bias of approximately 5% in New York and BKS (2016) estimate forecast bias of approximately 3-5% in Los Angeles. Experiments in which students are randomly assigned to classrooms also find little forecast bias (Kane and Staiger 2008, Kane et al. 2013). Even Rothstein's own analysis – setting aside our disagreements about methodology – yields estimates of forecast bias of approximately 5-15% across the specifications he explores. In sum, all of these studies imply that policies which select teachers on the basis of their estimated VA would achieve at least 85% of the gains one would expect if VA were unbiased.[28] The literature therefore paints a consistent picture showing that VA metrics provide useful information about teachers' impacts on students' outcomes.

At this point, our view is that resolving the exact amount of forecast bias in VA estimates is of secondary importance relative to other questions in the policy debate on VA measures. For example, studying how the properties of VA measures change when they are used in high-stakes settings and understanding how VA can be combined with other measures of teacher quality (such as classroom observations or principal ratings) to measure teachers' long-term impacts more accurately are critical issues that remain to be resolved in future research.

[28]Rothstein argues that the estimates from these studies differ to a larger degree by focusing on "teacher-level" bias instead of forecast bias. However, as we discussed in the introduction, all of the studies in question estimate only forecast bias, not teacher-level bias. Hence, we view forecast bias as the relevant measure for comparisons.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

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

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014b. "Measuring the Impact of Teachers II: The Long-Term Impacts of Teachers." *American Economic Review* 104(9): 2633-2679.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2016. "Using Lagged Outcomes to Evaluate Bias in Value-Added Models." *American Economic Review Papers and Proceedings* 106(5): 393-399.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593-1660.

Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3): 111-134.

Goldhaber, Dan, and Duncan D. Chaplin. 2015. "Assessing the "Rothstein Falsification Test": Does It Really Show Teacher Value-Added Models Are Biased?" *Journal of Research on Educational Effectiveness* 8(1): 8–34.

Kane, Thomas J., and Douglas O. Staiger. 2001. "Improving School Accountability Measures." NBER Working Paper 8156.

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

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

Kane, Thomas J., Douglas O. Staiger, and Andrew Bacher-Hicks. 2016. "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles." Unpublished manuscript, Harvard University.

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

Rothstein, Jesse. 2014. "Teacher Quality When Supply Matters." *American Economic Review* 105(1): 100-130.

Rothstein, Jesse. 2014. "Revisiting the Impacts of Teachers." UC-Berkeley Working Paper.

Rothstein, Jesse. 2016. "Revisiting the Impacts of Teachers." Forthcoming, *American Economic Review.*

Online Appendix A. Teacher Followers and Prior Scores

In this appendix, we provide further detail on why including teachers who follow students across grades can produce correlations between changes in current VA and prior test scores across cohorts.

Consider the relationship between changes in math test scores and math teacher VA from 1994 to 1995 in 5th grade in a given school. Suppose a teacher with high estimated VA leaves 5th grade after 1994 and is replaced by a teacher with average VA; assume there are no other changes in the teaching roster. We know that the high-VA teacher who departed did not teach the children who were in grade 5 in 1995 when they were in 4th grade in 1994 (because she taught 5th grade in 1994). However, she may have taught the children who were in grade 5 in 1994 when they were in 4th grade in 1993. As a result, the high VA of the departing teacher is positively correlated with lagged test scores of the cohort that reaches 5th grade in 1994, but not the test scores of those who reach 5th grade in 1995. This effect makes lagged (4th grade) test scores fall on average across the two cohorts. Since (by construction) teacher VA is also falling in this example, there is a positive correlation between changes in lagged (4th grade) scores across the two cohorts and mean teacher VA.

It is useful to distinguish between two separate channels that drive this correlation. The first channel is fluctuations in student test scores that are not related to the persistent component of teacher value-added, i.e., noise in student test scores. The teachers in 5th grade in 1994 could have higher estimated VA simply because her students in 4th grade test in 1993 did particularly well by chance (e.g., because the curriculum in the school happened to be well aligned with the test questions that year). This creates a mechanical correlation between lagged scores and VA estimates but has no bearing on our estimate of forecast bias using current test scores. Second, the correlation could be driven by teacher treatment effects. If the 5th grade teachers in 1994 were of truly high quality, they would affect the performance of 4th graders in 1993 (because some of them taught 4th grade in 1993), but not the 4th graders in 1994 (because we know they are teaching 5th grade in 1994). Note that, in contrast to the first channel, the direct treatment effect of teachers in prior grades could potentially bias our estimate of λ, as having better teachers in prior school years can increase current scores. The magnitude of bias depends upon the rate of fade-out in the sample where a teacher teaches the same child twice. The fact that the estimates of λ do not change when we exclude followers (Columns 1 and 2 of Table 4) shows that in practice, there is little or no bias in our estimate of λ from this latter channel.

Online Appendix B. Residualization Using Within-Teacher Variation

In a standard partial-regression implementation of a multivariate regression model, one must residualize *both* the left- and right-hand side variables with respect to the covariates to obtain a consistent estimate of the regression coefficient of interest. In this appendix, we show why one should not residualize the right hand size variable (teacher VA) with respect to covariates X_{it} when the effects of the covariates on long-term outcomes Y_{it}^* are estimated using *within-teacher* variation.

Suppose the statistical model for earnings is

$$Y_i^* = \alpha + \kappa m_j + \beta^Y X_{it} + \} \eta_{it}, \tag{1}$$

with η_{it} orthogonal to X_{it} and m_j defined as normalized teacher VA, as in equations (2) and (4) in CFR-II.

First, observe that if we knew β^Y, we could mechanically construct $Y_{it} = Y_i^* - \beta^Y X_{it}$ and then simply regress Y_{it} on m_j (without including any controls) to obtain an unbiased estimate of κ under the selection on observables assumption in CFR-II (Assumption 2). In this case, residualizing value-added m_j with respect to X_{it} would *not* yield a consistent estimate of κ because the true model is $Y_{it} = \alpha + \kappa m_j + \eta_{it}$.

Now suppose that we do not know β^Y and estimate it using an OLS regression (without teacher fixed effects) of the form

$$Y_i^* = a + b^Y X_{it} + \varepsilon_{it}.$$

Here b^Y does not provide a consistent estimate of β^Y if teacher VA m_j is correlated with X_{it}: b^Y converges to $\beta^Y + \kappa Cov(m_j, X_{it})/Var(X_{it})$. Since b^Y is not a consistent estimate of β^Y, one cannot simply regress $Y_i^* - b^Y X_{it}$ on m_j to obtain a consistent estimate of κ. Intuitively, the reason one must residualize both Y_{it} and m_j in a multivariate regression is that an OLS regression of Y_{it} on X_{it} does not produce a consistent estimate of the structural parameter β^Y in (1) because it partly picks up the effect of m_j, which is correlated with X_{it}. To correct for the incorrect estimate of β^Y, one must residualize the right-hand-side variable m_j with respect to X_{it} and then regress the earnings residuals Y_{it} on the VA residuals $\tilde{m}_{jt} = m_j - \gamma X_{it}$.

Now consider our approach, where we estimate β^Y using an OLS regression *with* teacher fixed effects a_j:

$$Y_i^* = a_j + b_f^Y X_{it} + \varepsilon_{it}^{'}.$$

Here, the coefficient b_f^Y is identified purely from within-teacher variation in X_{it} that is mechanically uncorrelated with variation in m_j. Therefore, under the model in (1), the coefficient b_f^Y converges

to β^Y and hence regressing $Y_i^* - b_f^Y X_{it}$ on m_j yields a consistent estimate of κ, for the same reason that regressing $Y_i^* - \beta^Y X_{it}$ on m_j when β^Y is known yields a consistent estimate of κ in the first case considered above. In contrast, using residual VA $\tilde{m}_{jt} = m_j - \gamma X_{it}$ in the second regression would yield an inconsistent estimate of κ. Intuitively, when we use within-teacher variation to estimate β^Y, we immediately obtain a consistent estimate of the effect of X on earnings that is not contaminated by the correlation with teacher value-added. Hence, one simply has to regress the outcome residual on VA to estimate the effect of teacher VA in the second stage.

Online Appendix C. Stata Code for Simulations

Imputation of Missing Data

```
1
2 ****************
3 * This simulation shows that imputing zeros reduces the coefficient in the
regression of changes in scores on changes in mean VA when VA is correlated
across teachers in a cell
4 ****************
5
6 clear all
7
8 *1 Generate data at the school-year-teacher level
9 set obs 1000000
10 set seed 5071788
11 g year = mod(_n-1,2)+1
12 g teacher = ceil(_n/2)
13 g school = ceil(teacher/2)
14
15 *2 Generate correlated VA within cells
16 global corr = 0.2
17 tsset teacher year
18 g va = rnormal(0,.1) if year == 1
19 replace va = (${corr}*l.va + sqrt(1-${corr}^2)*rnormal(0,.1)) if year == 2
20
21 *3 Generate missing data
22 g rand = runiform()
23 g miss = rand<.2
24 g va_miss = va
25 replace va_miss = . if miss==1
26
27 *4 Generate scores
28 g score = va + rnormal(0,1)
29 g score_miss = score
30 replace score_miss = . if va_miss == .
31
32 *5 Imputation of 0 for missing data
33 g va_impute = va_miss
34 replace va_impute = 0 if va_miss == .
35
36 *6 Collapse to school-year and run regressions
37 collapse va va_miss va_impute score score_miss, by(school year)
38 tsset school year
39
40 *7 Results
41 log using imputation.smcl, replace
42
43 _eststo clear
44 _eststo Full_Sample: reg d.score d.va // Coefficient in full sample
(ideal data)
45 _eststo No_Missing: reg d.score_miss d.va_miss // Coefficient on subsample
with no missing data
46 _eststo Impute_0s: reg d.score d.va_impute // Coefficient with imputation
```

```
is downward-biased
49 esttab _all, mtitles title("Missing Data Imputation Simulation Results")
se not
50 log close
```

Prior Test Scores (Table 3)

```
1 *This simulation shows that prior test score changes with be correlated
with changes in current mean VA across cohorts when there are school-year
level shocks
2 *Simulates data for one subject, so shocks should be interpreted as school-
year-subject shocks
3 *The program simulates class-level data, incorporating teacher effects,
class effects, student-level noise, and a school-year shock common to all
classrooms.
4
5 clear all
6 set seed 717806
7 set more off
8
9 * Parameters governing simulation
10 global min_grade = 3 // Minimum grade level (for readability)
11 global min_year = 1992 // Start year (for readability)
12 global n_school = 10000 // Number of schools
13 global n_year = 6 // Number of years
14 global n_grade = 6 // Number of grades taught per school
15 global n_rooms = 4 // Number of classrooms per school and grade
16 global n_class = 25 // Number of students per class
17 global var_tot = 0.25 // Total variance of scores
18 global sd_va = 0.10 // Standard deviation of value added
19 global sd_class_shock = 0.08 // Standard deviation of classroom-level
shocks
20 global sd_sy_shock = .08 // Standard deviation of school-by-year shocks
21 global rho_sy = 0.35 // Autocorrelation on school-by-year shocks
22
23 * Generate basic data
24 set obs `= ${n_school} * ${n_grade} * ${n_rooms} * ${n_year}'
25 g school = ceil(_n / (${n_grade} * ${n_rooms} * ${n_year}))
26 g grade = mod(ceil(_n / (${n_rooms} * ${n_year})) - 1 , ${n_grade}) +
${min_grade}
27 g teacher = ceil(_n / ${n_year})
28 g year = mod(_n - 1, ${n_year}) + ${min_year}
29 g id = rnormal()
30
31 * Replace some teachers in 1997
32 * Only in grades 5-8 (since others not used in experiment at end)
33 g replacement = mod(teacher , ${n_rooms}) < 1 if year == ${min_year}
34 replace replacement = replacement[_n-1] if year > ${min_year}
35 replace grade = grade - 1 if replace == 1 & year >= 1997
36 replace school = mod(school , ${n_school}) + 1 if grade ==
(${min_grade}+1) & year >= 1997 & replace == 1
37 replace grade = ${min_grade} + ${n_grade} - 1 if grade == (${min_grade}+1)
& year >= 1997 & replace == 1
38 replace grade = grade + 1 if replace == 1 & year >= 1997 & grade < (
```

```
${min_grade}+1)
39
40 * Generate true VA and class shocks
41 g class_shock = rnormal(0, ${sd_class_shock})
42 g va_true = rnormal(0,${sd_va}) if year==${min_year}
43 replace va_true = va_true[_n-1] if va_true==.
44
45 * Generate average lagged true VA and average lagged class shocks as
"double lags" of true
VA and class shocks
46 sort school year grade id
47 g temp1 = va_true[_n - ${n_rooms} * (${n_grade} + 1)] if year >
${min_year} & grade > ${min_grade}
48 g temp2 = temp1[_n - ${n_rooms} * (${n_grade} + 1)] if year > ${min_year}
& grade > ${min_grade}
49 g temp3 = class_shock[_n - ${n_rooms} * (${n_grade} + 1)] if year >
${min_year} & grade > ${min_grade}
50 g temp4 = temp3[_n - ${n_rooms} * (${n_grade} + 1)] if year > ${min_year}
& grade > ${min_grade}
51 by school year grade: egen l_va_true = mean(temp1)
52 by school year grade: egen l2_va_true = mean(temp2)
53 by school year grade: egen l_class_shock = mean(temp3)
54 by school year grade: egen l2_class_shock = mean(temp4)
55 drop temp*
56
57 * Generate school-by-year shocks
58 g sy_shock = rnormal(0 , ${sd_sy_shock} * sqrt(1 - ${rho_sy}^2)) if mod(_n
,${n_rooms} * ${n_grade}) == 1
59 replace sy_shock = sy_shock / sqrt(1 - ${rho_sy}^2) if year ==
${min_year} & ~missing(sy_shock)
60 replace sy_shock = sy_shock[_n - 1] if missing(
sy_shock)
61 replace sy_shock = sy_shock + ${rho_sy} * sy_shock[_n - ${n_rooms} *
${n_grade}] if year >
${min_year}
62 g l_sy_shock = sy_shock[_n - ${n_rooms} * ${n_grade}] if year >
${min_year}
63 g l2_sy_shock = sy_shock[_n - 2 * ${n_rooms} * ${n_grade}] if year > (
${min_year} + 1)
64
65 * Generate classroom average score, lagged, and twice-lagged scores
66 global sd_indv = sqrt((${var_tot} - ${sd_va}^2 - ${sd_sy_shock}^2 -
${sd_class_shock}^2) /
${n_class})
67 g l2_score = (l2_va_true + l2_sy_shock + l2_class_shock +
rnormal(0,${sd_indv}))
68 g l_score = l_va_true + l_sy_shock + l_class_shock + rnormal(0,${sd_indv})
69 g score = va_true + sy_shock + class_shock + rnormal(0,${sd_indv})
70
71 *Make dataset balanced panel
72 replace l_score = . if l2_score == .
73 replace score = . if l_score == .
74
75 * Residualize scores using a single lag
76 sort teacher year
77 sum score
78 global tot_var = r(Var)
```

```
tsset teacher year
corr score l.score, c
global teach_var = r(cov_12)
global ind_var = (${n_class} / (${n_class} - 1)) * (${var_tot} -
${tot_var})
global class_var = ${var_tot} - ${ind_var} - ${teach_var}

* Construct leave-two-out shrinkage and VA estimate
g temp = ~inrange(year , 1996 , 1997) & ~missing(score)
by teacher: egen temp1 = sum(temp)
g shrinkage = ${teach_var} / (${teach_var} + ${class_var} / temp1 +
${ind_var} / (${n_class}
* temp1))
g temp2 = score if ~inrange(year , 1996 , 1997) & ~missing(score)
by teacher: egen temp3 = mean(temp2)
g va = temp3 * shrinkage if inrange(year,1996,1997)
drop temp*

* Construct Rothstein (2016) leave-three-out shrinkage and VA estimate
g temp = ~inrange(year , 1995 , 1997) & ~missing(score)
by teacher: egen temp1 = sum(temp)
g shrinkage_3out = ${teach_var} / (${teach_var} + ${class_var} / temp1 +
${ind_var} / (
${n_class} * temp1))
g temp2 = score if ~inrange(year , 1995 , 1997) & ~missing(score)
by teacher: egen temp3 = mean(temp2)
g va_3out = temp3 * shrinkage_3out if inrange(year,1995,1997)
drop temp*

* Construct prior shrinkage and VA estimate
g temp = (year < 1997) & ~missing(score)
bys teacher: egen temp1 = sum(temp)
g shrinkage_prior = ${teach_var} / (${teach_var} + ${class_var} / temp1 +
${ind_var} / (
${n_class} * temp1))
g temp2 = score if (year < 1997) & ~missing(score)
bys teacher: egen temp3 = mean(temp2)
drop temp*

save vam_simulation, replace

* Collapse data to school-grade-year level to implement quasi-
experimental analysis
keep if inrange(year,1996,1997)
collapse score l_score va va_3out va_true, by(school grade year)
egen sy = group(school year)
egen sg = group(school grade)
tsset sg year
save vam_simulation_collapse, replace

* Results
log using lagged_score_simulation.smcl, replace
eststo clear
_eststo d_score: qui reg d.score d.va
_eststo d_score_sy: qui reg d.score d.va , a(sy)
_eststo d_score_3out: qui reg d.score d.va_3out
_eststo lag_d_score: qui reg d.l_score d.va
```

```
128 _eststo lag_d_score_sy: qui reg d.l_score d.va , a(sy)
129 _eststo lag_d_score_3out: qui reg d.l_score d.va_3out
130 esttab _all, mtitles title("Quasi-Experimental Forecast Bias Estimates")
se not
131 log close
```

Long-Term Effects of VA (Table 5)

```
1 *This simulation shows that estimates of long-term impacts are downward-
biased in a multi-variable regression because VA is estimated with error and
is correlated with X
2
3 clear all
4
5 *Specify target: true effect of 1 unit increase in va_score on earnings
6 *Note that this is equivalent to effect of 1 unit increase in mu (not m =
mu/sd(mu))
7 global true_coeff = 100
8
9 ************************PART 1*********************
10 ********************Generate data*******************
11 ****************************************************
12
13 set obs 1000000
14 global n_class = 20
15 global classes_per_teach = 10
16 g class = ceil(_n/$n_class)
17 g teacher = ceil(_n/(${n_class}*${classes_per_teach}))
18
19 *Generate test-score VA (mu_j)
20 bys teacher: g temp = _n
21 g temp1 = rnormal(0,0.1) if temp == 1
22 bys teacher: egen va_score = mean(temp1)
23
24 *Generate pure earnings component of VA
25 g temp2 = rnormal(0,0.1) if temp == 1
26 bys teacher: egen va_earn = mean(temp2)
27 drop temp*
28
29 *Generate total earnings VA (tau_j)
30 g va_comb = va_score + va_earn
31
32 *Generate covariate X correlated with teacher's total earnings VA
33 global rho = 0.33
34 g x = (${rho}*va_comb + (1-${rho})*rnormal(0,0.1))/sqrt(${rho}^2+(1-
${rho})^2)
35
36 *Generate scores and earnings
37 *Note that only va_score affects test scores, while both va_score and
va_earn affect earnings
38 g score = va_score + x + rnormal(0,sqrt(1-.1^2))
39 g earn = ${true_coeff}*va_comb + 10*x + rnormal(0,10)
40
41 ***********************PART 2. ********************
42 **Estimate VA using within-teacher residualization**
```

```
43 ************************************************
44
45 * Residualize scores using within teacher variation as in CFR (2014b)
46 qui areg score x, a(teacher)
47 predict score_res, dr
48
49 * Estimate teacher-level variance
50 preserve
51 collapse score_res, by(teacher class)
52 tsset teacher class
53 qui corr score_res l.score_res, c
54 global teach_var = r(cov_12)
55 restore
56
57 * Estimate residual variance and shrinkage
58 sum score_res
59 global tot_var = r(Var)
60 global ind_var = ${tot_var} - ${teach_var}
61 scalar shrinkage = ${teach_var}/(${teach_var} + ${ind_var}/(${n_class} * (
${classes_per_teach}-1)))
62
63 * Estimate Leave-Out VA
64 bys teacher: egen temp = mean(score_res)
65 bys teacher class: egen temp1 = mean(score_res)
66 g va = (${classes_per_teach}*temp - temp1)/(${classes_per_teach}-1)
*shrinkage
67 drop temp*
68
long_term_controls_simulation - Printed on 2/4/2015 9:33:20 PM
Page 2
69 *Confirm that regressing test score residuals on VA gives a coeff of 1
70 reg score_res va
71
72 **********PART 3**************************
73 **Alternative Estimators of Teachers' Long-Term Effects**
74 ************************************************
75
76 log using long_term_controls_simulation.smcl, replace
77
78 ***Column 1. Estimate long-term effects using two-step residualization as
in CFR (2014b)
79 *Yields correct estimate of long-run effects as expected
80 qui areg earn x, a(teacher)
81 predict earn_res, dr
82 reg earn_res va, cl(teacher)
83
84 ***Column 2. Estimate long-term effects using multivariable regression as
in Rothstein (2016)
85 *Yields attenuated coefficient as expected
86 reg earn va x, cl(teacher)
87
88 *Column 3: Rothstein (2014) 2SLS estimator yields estimate similar to OLS
in #3
89 ivreg earn (score_res = va) x
90
91 *Note: This is because first stage coef is very close to 1
92 reg score_res va x
```

```
93
94 *Column 4: Rothstein (2014) OLS multivariable estimator yields correct
estimate if we are able to control for (unobserved) true earnings VA
95 reg earn va_comb x
96
97 log close
```

TABLE 1

Effects of Imputing Missing VA: Empirical Results

	Full Sample Excluding Teachers with Missing VA	Sch-Gr-Subj. Cells with No Missing Data	Cells with < 25% Missing Data, Imputing 0's to Missing Data	Full Sample, Imputing 0's to Missing Data
	Dep. Var.: Change in Mean Score Across Cohorts			
	(1)	(2)	(3)	(4)
	Panel A: CFR (2014a) New York Sample			
Change in Mean VA across Cohorts	0.974	0.990	0.952	0.877
	(0.033)	(0.045)	(0.032)	(0.026)
Grades	4 to 8	4 to 8	4 to 8	4 to 8
N. Sch. x Grades x Subject x Year Cells	59,770	17,859	38,958	62,209
Percent of Obs. With Non-Imputed VA	100.0	100.0	93.8	83.6
	Panel B: Rothstein (2016) North Carolina Sample			
Change in Mean VA across Cohorts	1.097	1.081	1.100	0.936
	(0.022)	(0.043)	(0.035)	(0.022)
Grades	3 to 5	3 to 5	3 to 5	3 to 5
N. Sch. x Grade x Subject x Year Cells	79,466	23,445	34,495	91,221
Percent of Obs. With Non-Imputed VA	100.0	100.0	94.4	72.6
	Panel C: Bacher-Hicks, Kane, Staiger (2016) LAUSD Sample			
Change in Mean VA across Cohorts	1.030	0.973		0.993
	(0.044)	(0.048)		(0.049)
Grades	4 to 8	4 to 8		4 to 8
N. Sch. x Grade x Subject x Year Cells	14,186	8,974		14,292
Percent of Obs. With Non-Imputed VA	100.0	100.0		92.0

Notes: This table presents estimates from regressions of the change in mean test scores across consecutive cohorts within a school-grade-subject cell on changes in mean teacher value-added (VA). Panel A replicates estimates reported in CFR-I (2014, Tables 4 and 5), while Panel B replicates results from Rothstein (2016, Appendix Tables A4 and A5), and Panel C replicates results from Bacher-Hicks, Kane, and Staiger (2016, Table 3). Column 1 restricts the sample to students with non-missing teacher VA. Column 2 restricts the sample to school-grade-year cells with no missing teacher VA. Column 3 restricts the sample to school-grade-year cells where VA is missing for less than 25% of the observations, imputing VA of 0 to teachers with missing VA. Column 4 uses the full sample, imputing VA of 0 to teachers with missing VA. All specifications include year fixed effects. See notes to Tables 4 and 5 of CFR-I for further details on the specifications.

TABLE 2a

Stylized Example: Correlation Between Changes in Prior Scores and Teacher VA with Independent Shocks

	Year				
	1992	1993	1994	1995	1996
School-wide Math Test Score Shock:	**0**	**+1**	0	0	0
Cohort 1 Grade Level (Math Test Score):	3rd (0)	4th (+1)	5th (0)	6th (0)	7th (0)
Cohort 2 Grade Level (Math Test Score):	2nd (0)	3rd (+1)	4th (0)	5th (0)	6th (0)
VA of Math Teacher who leaves 5th grade after 1994:			+1		
VA of Math Teacher who enters 5th grade in 1995:				0	
Cross-Cohort Change in 5th Grade Teacher VA ('95 - 94)			-1		
Cross-Cohort Change in Lagged/4th Grade Scores ('94 - '93)			-1		
Cross-Cohort Change in 5th Grade Scores ('95 - '94)			0		

Notes: This table illustrates how a school-level shock to test scores can create a spurious positive correlation between changes in students' lagged scores and changes in measured teacher VA. We assume VA is measured using data from 1993 (for the departing teacher) and 1996 (for the entering teacher) in this example. The table shows the effects of a non-persistent shock to 1993 math test scores of +1 unit in a school. This shock raises 4th grade scores for Cohort 1 relative to Cohort 2, while also raising measured VA for the 5th grade teacher who leaves after 1994 relative to the teacher who enters 5th grade in 1995. This means that the cross-cohort changes in lagged scores and measured VA will both be negative, i.e. that there is a positive correlation between changes in mean VA and changes in prior scores.

TABLE 2b

Stylized Example: Correlation Between Changes in Prior Scores and Leave-3-Out VA with Serially Correlated Shocks

	Year				
	1992	1993	1994	1995	1996
School-wide Math Test Score Shock:	**+1**	**+0.5**	0	0	0
Cohort 1 Grade Level (Math Test Score):	3rd (+1)	4th (+0.5)	5th (0)	6th (0)	7th (0)
Cohort 2 Grade Level (Math Test Score):	2nd (+1)	3rd (+0.5)	4th (0)	5th (0)	6th (0)
VA of Math Teacher who leaves 5th grade after 1994:			+1		
VA of Math Teacher who enters 5th grade in 1995:				0	
Cross-Cohort Change in 5th Grade Teacher VA ('95 - 94)			-1		
Cross-Cohort Change in Lagged/4th Grade Scores ('94 - '93)			-0.5		
Cross-Cohort Change in 5th Grade Scores ('95 - '94)			0		

Notes: This table illustrates how a serially correlated shock to test scores can create a spurious positive correlation between changes in students' scores and changes in measured teacher VA, even if we apply a "leave 3 out" rule when calculating teacher VA, as in Rothstein (2016, Appendix B). We assume VA is measured using data from 1992 (for the departing teacher) and 1996 (for the entering teacher) in this example. The table shows the effects of a shock to 1992 math test scores within a school that decays over time. This shock raises 4th grade scores for Cohort 1 relative to Cohort 2. Using the "leave 3 out" rule, VA for the departing teacher in 1994 is measured using data from 1992, while VA for the entering teacher is measured using data from 1996. Thus measured VA for the 5th grade teacher who leaves after 1994 will be higher relative to the teacher who enters 5th grade in 1995. This means that the cross-cohort changes in lagged scores and measured VA will both be negative, i.e. that there is a positive correlation between changes in mean VA and changes in prior scores.

TABLE 3

Prior Test Score "Placebo" Tests: Simulation Results

	CFR- I Baseline (1)	Leave-Three-Out VA (2)	Sch.-Year (Subj.) Fixed Effects (3)
Panel A. Dependent Variable: Δ Mean Current Score			
Change in Mean VA	0.997	0.991	0.972
	(0.024)	(0.029)	(0.018)
Panel B. Dependent Variable: Δ Mean Lagged Score			
Change in Mean VA	0.138	0.097	0.009
	(0.023)	(0.028)	(0.017)

Notes: This table presents estimates using data that is simulated from a model with school by year shocks to test scores. Each cell of this table reports estimates from a separate regression, with standard errors in parentheses. The dependent variable in Panel A is the change in current scores, while the dependent variable in Panel B is the change in lagged scores. Column 1 is the baseline specification using in CFR-I. Column 2 reports estimates where current VA is calculated excluding three calendar years of data. Column 3 includes school-year fixed effects (there is no variable for subject in our simulation).

TABLE 4

Prior Test Score "Placebo" Tests: Empirical Results

	Baseline	No Followers	Between-School (Excludes Followers)	Within-School (Includes Followers)	No Followers w/ Sch.-Year-Subj. FE's
	(1)	(2)	(3)	(4)	(5)
Panel A. Dependent Variable: Δ Mean Current Score					
Change in Mean VA (λ)	0.957	0.923	0.968	0.950	0.942
	(0.034)	(0.037)	(0.069)	(0.051)	(0.042)
P-value: Coeff = Baseline		[0.361]	[0.870]	[0.890]	[0.722]
Panel B. Dependent Variable: Δ Mean Lagged Score					
Change in Mean VA (θ)	0.171	0.052	0.031	0.258	-0.023
	(0.034)	(0.036)	(0.068)	(0.050)	(0.043)
P-value: Coeff = 0	[0.000]	[0.145]	[0.650]	[0.000]	[0.596]
Panel C. Predicted Values of λ Assuming ψ = 0.66 as in Rothstein (2016)					
Change in VA		0.879	0.865	1.015	0.829

Notes: Each cell in Panels A and B of this table reports estimates from a separate regression, with the coefficient in the top row, standard error (in parenthesis) in the following row, and p-value [in brackets] in the third row. The dependent variable in Panel A is the change in mean current scores across cohorts, while the dependent variable in row 2 is the change in mean lagged scores across cohorts. All regressions in Columns 1 through 4 include school-year fixed effects. Column 1 is the baseline specification used in CFR-I and Rothstein (2016, Table 2, Col. 1). Column 2 reports 2SLS estimates, instrumenting for the change in mean VA with the changes in mean VA excluding teachers who switch from the previous grade to the current grade. Column 3 reports 2SLS estimates, instrumenting for the change in mean VA with the changes in mean VA excluding teachers who switched grades within the school. Column 4 reports 2SLS estimates, instrumenting for the change in mean VA with the changes in mean VA from teachers who switched grades within the school. Column 5 replicates 2 including school-year-subject fixed effects. See notes to Table 4 in CFR-I for further details on these specifications. Panel C reports predicted values of the coefficient in Panel A under the assumption that a 1 unit increase in prior scores increases current scores by 0.66, starting from the baseline values in Column 1. See Section II.D for further details on this panel.

TABLE 5

Long-Term Effects of Teacher VA: Simulation Results

Estimation Method:	CFR Two-Step Estimator	Rothstein Multivariable OLS	Rothstein Multivariable 2SLS
	Dependent Variable: Student Earnings ($)		
	(1)	(2)	(3)
Teacher Test-Score VA (μ)	99.31	74.84	79.51
	(2.006)	(1.542)	(1.121)
Covariate (X)		65.86	54.34
		(0.839)	(0.842)

Notes: This table reports results from simulated data in which student earnings are a function of teacher VA and a correlated covariate X. The true effect of teacher VA on student earnings is $100 and the true effect of the covariate X is $10. Each column reports regression estimates using a different estimator. Column 1 shows estimates obtained from the two-step estimator used in CFR-II. Column 2 shows estimates obtained from Rothstein's (2016) multivariable OLS regression. Column 3 shows estimates from Rothstein's (2014) 2SLS estimator.