Centre for Economic Performance

Discussion Paper

ISSN 2042-2695

No. 2214
September 2026

Revisiting test score forecasts and teacher value-added

Tom Ahn
Esteban M. Aucejo
Jonathan James

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UKRI Economic and Social Research Council

Abstract
We develop a framework for comparing forecasts in teacher value-added models. We apply it to interpret the teacher-switching test of Chetty, Friedman, and Rockoff (2014), showing that the teacher-switching coefficient cannot be mechanically interpreted as a test of forecast unbiasedness: the correct null may differ from one, and a coefficient of one is insufficient for forecast unbiasedness. We also analyze forecast choice for ranking teachers and estimating teachers' effects on long-term outcomes. Although forecasts including current-year data may contain contemporaneous classroom shocks, they contain more teacher signal than leave-year-out forecasts, which improves rankings and may increase precision in long-term-outcome regressions.

Keywords: teacher value added; forecast construction; forecast aggregation; forecast unbiasedness; teacher rankings
JEL codes: I21; J24; C53; C52

This paper was produced as part of the Education Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Tom Ahn, Naval Postgraduate School. Esteban M. Aucejo, Arizona State University and Centre for Economic Performance at the London School of Economics. Jonathan James, California Polytechnic State University.

Published by
Centre for Economic Performance
London School of Economic and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

Forecasts produced by teacher value-added models play a central role in education research and policy (Hanushek and Rivkin, 2012). The goal of these forecasts is to isolate a teacher's contribution to observed outcomes by pooling information across multiple students. In practice, however, observed outcomes also reflect factors unrelated to the teacher's contribution, which can affect the forecast. The concern that forecasts may be contaminated by transitory or persistent non-teacher inputs has received significant attention in the literature (Koedel et al., 2015).

Prior work has taken two main approaches to non-teacher inputs. For classroom shocks, or transitory non-teacher inputs, a common approach following Chetty et al. (2014a) is to construct leave-year-out forecasts, which exclude all contemporaneous classroom data and therefore prevent contemporaneous classroom shocks from entering the forecast. However, there has been little discussion of whether the benefit of removing these shocks outweighs the cost of discarding additional teacher signal. The appropriate forecast choice likely depends on how the forecasts are ultimately used and may therefore differ across applications.

Addressing persistent non-teacher inputs is more difficult because they may remain in the forecast even after contemporaneous classroom shocks are removed. To assess whether persistent non-teacher inputs enter value-added forecasts, Chetty et al. (2014a) propose a teacher-switching test, which has become a standard test in this literature. In the teacher-switching test, a coefficient of one is used as evidence that the forecasts are unbiased. However, the interpretation of the teacher-switching coefficient remains debated (Rothstein, 2017; Chetty et al., 2017).

The goal of this paper is to develop a framework for understanding and comparing forecasts in teacher value-added models. For a given forecast, our framework decomposes the forecast into the contributions of teacher value added, classroom shocks, and idiosyncratic student-level shocks. We use this framework for two distinct purposes. First, the framework makes clear which sources of test-score variation enter the forecast, providing a direct way to interpret coefficients from regressions such as the teacher-switching test in Chetty et al. (2014a). Second, the framework shows how the data used to construct the forecast affect the overall signal strength of the test-score components and identifies potential sources of forecast bias. This provides a way to compare the trade-offs of different forecasting choices in policy applications.

We make three contributions. First, we examine forecast calibration and the interpretation of forecast-based regression tests. In a forecast calibration regression, test-score outcomes are regressed on model-implied forecasts to determine whether the forecasts are cor-

rectly calibrated to the outcome. Forecast-based regression tests have the same regression form, but are often used to learn about features of the test-score process or the forecasts themselves. We examine tests of correct forecast scale (Chetty et al., 2014a), within-year variation in test scores (Chetty et al., 2011), and across-teacher information in test scores (Rothstein, 2017).[1] For each test, the framework shows what determines the test coefficient and when the test has a direct interpretation.

Using these results, we analyze across-teacher information within elementary schools in North Carolina. As discussed by Rothstein (2017), averaged teacher-level forecasts may be incorrectly calibrated to average school-level test scores if there is across-teacher information in any of the test-score components within schools. We show that across-teacher information has a substantial effect on aggregated forecasts: the correct aggregate forecast weight is 18.8% to 33.7% larger than the weight implied by averaged teacher-level forecasts. This indicates that analyses that rely on correctly calibrated aggregate forecasts should not use averaged teacher-level forecasts that ignore across-teacher information.

Our second contribution applies the framework to the widely used teacher-switching test proposed by Chetty et al. (2014a). This forecast-based regression test is designed to detect bias in teacher-level value-added forecasts arising from persistent non-teacher inputs. The test relates year-over-year changes in average school-grade test scores to changes in average value-added forecasts caused by turnover among teachers assigned to the school-grade. Chetty et al. (2014a) argue that a coefficient of one in the teacher-switching regression indicates that the teacher-level value-added forecasts are unbiased.

We show that the teacher-switching coefficient cannot be interpreted mechanically as a test of teacher-level forecast bias for two reasons. First, because the test relies on forecast aggregation, across-teacher information can cause the correct null value of the coefficient to differ from one, even when teacher-level forecasts are unbiased for teacher value added. Thus, the test may reject even when teacher-level forecasts are unbiased. Second, a coefficient of one is not sufficient for forecast unbiasedness, since offsetting effects can produce a coefficient of one even when the forecasts are biased.[2] Thus, failing to reject a coefficient of one, even with a precise estimate, cannot be interpreted as evidence that the teacher-level forecasts are unbiased. Interpreting the teacher-switching coefficient as a test of forecast unbiasedness

[1] In this paper, across-teacher information refers to information about one teacher's test-score components contained in the test-score data of other teachers. The relevant variation depends on what is being forecast. For example, aggregate school-grade forecasts would depend on across-teacher information within the school-grade, while year-over-year differences in aggregate school-grade forecasts would depend on across-teacher information shared by entering and exiting teachers within the school-grade.

[2] Rothstein (2017) critiques the teacher-switching test by arguing that a coefficient of one does not distinguish forecast unbiasedness from teacher-level unbiasedness. Our result is stronger, showing a coefficient of one can arise even when the forecasts are biased.

therefore requires additional restrictions or assumptions.

Our third contribution uses the framework to study the effects of forecast choice in two common policy applications: ranking teachers and estimating the effects of teachers on long-term student outcomes. Specifically, we consider the trade-off between leave-year-out forecasts and forecasts that additionally incorporate current-year data. For teacher ranking, the goal is to order teachers in a way that is closely aligned with their underlying value added. In this setting, our results show that the signal lost from using leave-year-out forecasts greatly outweighs the potential benefit of excluding contemporaneous classroom shocks. Forecasting approaches that incorporate more data produce more accurate rankings, especially in short panels where the loss of teacher signal from excluding data is largest.

We also compare these forecasts when they are used to estimate the effects of teachers on long-term student outcomes. In this setting, leave-year-out forecasts may reduce bias by excluding contemporaneous classroom shocks, but they can do so at the cost of lower precision. We propose a Hausman-type test that uses the comparison between leave-year-out and more inclusive forecasts to choose between forecasting approaches.

2 Framework

2.1 Basic Model

We begin with a standard value-added framework following Chetty et al. (2014a). In this model, the test score of student i, assigned to teacher j in year t, is given by

$$A_{it}^{*} = \beta X_{it} + \mu_{jt} + \theta_{jt} + \varepsilon_{it},$$

where X_{it} is a vector of observable student and classroom characteristics, μ_{jt} is teacher j's value added in year t, θ_{jt} is a class-level shock where we assume that each teacher teaches a single class per year, and ε_{it} is an idiosyncratic student-level shock.

Estimation of teacher value added is based on test score residuals. Given an unbiased estimate $\hat{\beta}$ of β, the residual for student i in year t is

$$A_{it} = A_{it}^{*} - \hat{\beta} X_{it} = \mu_{jt} + \theta_{jt} + \varepsilon_{it}. \tag{1}$$

While these moments could more generally vary over time, to simplify the derivations of our main results, we assume $\text{Var}(\mu_{jt}) = \sigma_{\mu}^{2}$ for all t and $\text{Cov}(\mu_{jt}, \mu_{jt'}) = \gamma_{\mu}$ for all $t \neq t'$. Under these assumptions, $\sigma_{\mu}^{2} - \gamma_{\mu}$ represents the variance of the transitory component of value added. The variances of the classroom and student-level shocks are denoted by σ_{θ}^{2} and

σ^2_ε, respectively.

2.2 Forecasting Framework

To compare different forecasting approaches, we place them in a common framework as forecasts of test-score residuals constructed from different information sets. The information set determines which components of test-score variation are treated as signal or noise, leading to different forecasts. We consider forecasts of A_{it} in Eq. (1) based on an information set $M^{(k)}_{it}$, which we assume is scalar to simplify the derivations, although more generally it may be a vector. Under the assumption that the test-score components are mean zero, the forecast is defined as the best linear predictor given the information set:

$$\hat{A}^{(k)}_{it} = \frac{\text{Cov}(A_{it}, M^{(k)}_{it})}{\text{Var}(M^{(k)}_{it})} M^{(k)}_{it} = w^{(k)} M^{(k)}_{it}.$$

Because test scores are linear in their components, the forecast can be written as the sum of the best linear predictors of the individual test-score components:

$$\begin{aligned} \hat{A}^{(k)}_{it} &= \underbrace{\left(\frac{\text{Cov}\left(\mu_{jt}, M^{(k)}_{it}\right)}{\text{Var}\left(M^{(k)}_{it}\right)} \right) M^{(k)}_{it}}_{\hat{\mu}^{(k)}_{jt}} + \underbrace{\left(\frac{\text{Cov}\left(\theta_{jt}, M^{(k)}_{it}\right)}{\text{Var}\left(M^{(k)}_{it}\right)} \right) M^{(k)}_{it}}_{\hat{\theta}^{(k)}_{jt}} + \underbrace{\left(\frac{\text{Cov}\left(\varepsilon_{it}, M^{(k)}_{it}\right)}{\text{Var}\left(M^{(k)}_{it}\right)} \right) M^{(k)}_{it}}_{\hat{\varepsilon}^{(k)}_{it}} \\ &= \left[w^{(k)}_\mu + w^{(k)}_\theta + w^{(k)}_\varepsilon \right] M^{(k)}_{it} \\ &= w^{(k)} M^{(k)}_{it}. \end{aligned} \tag{2}$$

This shows that the forecast weight, $w^{(k)}$, can be decomposed into individual test-score component weights, where $w^{(k)}_\mu$, $w^{(k)}_\theta$, and $w^{(k)}_\varepsilon$ denote the weights placed on the teacher, classroom, and idiosyncratic components of test-score variation, respectively. The magnitude of each component weight is determined by how strongly the information set correlates with that component.

We compare forecasts based on three different information sets: full-sample (FS), leave-one-out (LOO), and leave-year-out (LYO). Each information set is constructed from average test-score residuals. Define the average residual for teacher j in year t as

$$\bar{A}_{jt} \equiv \frac{1}{n} \sum_{i=1}^{n} A_{it},$$

and the leave-one-out average residual, excluding student i, as

$$\bar{A}_{jt}^{(-i)} \equiv \frac{1}{n-1} \sum_{i' \neq i} A_{i't}.$$

The information sets used in the analysis are defined as follows.[3]

$$\begin{aligned} M_j^{\text{FS}} &\equiv \frac{1}{T} \sum_{t=1}^{T} \bar{A}_{jt}, \\ M_{j,it}^{\text{LOO}} &\equiv \frac{1}{T} \left(\bar{A}_{jt}^{(-i)} + \sum_{t' \neq t} \bar{A}_{jt'} \right), \\ M_{j,t}^{\text{LYO}} &\equiv \frac{1}{T-1} \sum_{t' \neq t} \bar{A}_{jt'}. \end{aligned}$$

The main difference between these information sets is that the leave-year-out information set relies solely on across-year variation, while the full-sample and leave-one-out information sets are more inclusive because they also incorporate current-year variation.

Table 1 decomposes this forecast weight into the component weights $w_\mu^{(k)}$, $w_\theta^{(k)}$, and $w_\varepsilon^{(k)}$ for each information set. These weights show how each forecast loads on teacher value added, classroom shocks, and student-level noise. They characterize which sources of test-score variation enter the forecast and how much each component contributes to the overall prediction. The final column reports the total forecast weight, $w^{(k)}$. Its numerator represents the variation in the information set that is treated as test-score signal, while its denominator is the total variation in the information set. For example, the full-sample forecast treats all variation in the information set as signal, so the total forecast weight equals one.

The expressions in Table 1 provide useful insights for comparing forecasting approaches. Because the leave-year-out forecast excludes contemporaneous data, it is orthogonal to both contemporaneous classroom shocks and student-level shocks. This makes the leave-year-out test-score forecast a direct value-added forecast:

$$\frac{\text{Cov}\left(\mu_{jt}, \hat{A}_{it}^{(\text{LYO})}\right)}{\text{Var}\left(\hat{A}_{it}^{(\text{LYO})}\right)} = \frac{\text{Cov}\left(A_{it}, \hat{A}_{it}^{(\text{LYO})}\right)}{\text{Var}\left(\hat{A}_{it}^{(\text{LYO})}\right)} = 1.$$

This is the main appeal of leave-year-out predictors. More inclusive forecasts, by contrast, use contemporaneous data and therefore may load on classroom and student-level shocks

[3]To ease the comparisons, we use an equal-year-weighted leave-one-out average in LOO instead of an equal-student-weighted average. Either could be used, but the latter requires more notation.

Table 1: Forecast Weights by Information Set

Information set $M^{(k)}$	$w_\mu^{(k)}$	$w_\theta^{(k)}$	$w_\varepsilon^{(k)}$	$w^{(k)} = w_\mu^{(k)} + w_\theta^{(k)} + w_\varepsilon^{(k)}$
		Within- and Across-Year Variation		
Full-sample (M_j^{FS})	$\dfrac{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T}}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T} + \frac{\sigma_\varepsilon^2}{nT}}$	$\dfrac{\frac{\sigma_\theta^2}{T}}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T} + \frac{\sigma_\varepsilon^2}{nT}}$	$\dfrac{\frac{\sigma_\varepsilon^2}{nT}}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T} + \frac{\sigma_\varepsilon^2}{nT}}$	1
Leave-one-out ($M_{j,it}^{\text{LOO}}$)	$\dfrac{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T}}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T} + \frac{\sigma_\varepsilon^2}{\tilde{n}T}}$	$\dfrac{\frac{\sigma_\theta^2}{T}}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T} + \frac{\sigma_\varepsilon^2}{\tilde{n}T}}$	0	$\dfrac{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T}}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T} + \frac{\sigma_\varepsilon^2}{\tilde{n}T}}$
		Across-Year Variation Only		
Leave-year-out ($M_{j,t}^{\text{LYO}}$)	$\dfrac{\gamma_\mu}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T-1} + \frac{\sigma_\theta^2}{T-1} + \frac{\sigma_\varepsilon^2}{n(T-1)}}$	0	0	$\dfrac{\gamma_\mu}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T-1} + \frac{\sigma_\theta^2}{T-1} + \frac{\sigma_\varepsilon^2}{n(T-1)}}$

Note: The table reports the component weights in the best linear predictor of A_{it} for each information set, where $w^{(k)} = w_\mu^{(k)} + w_\theta^{(k)} + w_\varepsilon^{(k)}$. The parameters σ_μ^2, σ_θ^2, and σ_ε^2 denote the variances of teacher value added, classroom shocks, and idiosyncratic student shocks, respectively, and γ_μ denotes the across-year covariance in teacher value added. Class size is fixed at n and the panel length is T. For the leave-one-out forecast, $\tilde{n} = n\left(\dfrac{n-1}{n-1+\frac{1}{T}}\right) < n$ adjusts the effective class size after removing student i from the current-year classroom average.

that are not attributable to the teacher. This concern arises because, when teachers teach one classroom per year, within-year teacher value added cannot be separately identified from the contemporaneous classroom shock.

Next, we use this decomposition framework for two separate purposes. First, because the decomposition shows how test-score variation in the information set directly affects the forecast, Sections 3 and 4 use the framework to analyze coefficients in forecast-based regression tests. The framework shows when these coefficients have a direct interpretation and when they do not. Second, Section 5 uses the framework to study how forecast choice affects education policy analysis. Specifically, we examine the trade-off, across different policy applications, between using leave-year-out forecasts that remove current-year data and forecasts that retain current-year data. While leave-year-out forecasts avoid bias from contemporaneous classroom shocks, they also contain less teacher signal than forecasts that incorporate current-year data, making some forecasts more suitable than others depending on the type of analysis.

3 Interpreting Forecast-Based Regression Tests

Forecasts constructed from value-added models are often used in regression tests. These tests often resemble the forecast calibration regressions in Mincer and Zarnowitz (1969), comparing realized test-score residuals to model-derived forecasts. In value-added settings, however, forecast-based regression tests are often used for more than assessing whether a forecast is correctly calibrated to a particular outcome. Researchers also use these regressions to test particular features of the test-score components or of the forecasts themselves. Interpreting these tests requires examining how the forecasts used in the regressions are constructed and how their information sets relate to the outcome being predicted. In this section, we use the framework developed in Section 2 to study several forecast-based regression tests, showing what their coefficients recover and what can be inferred from their results. We begin with forecast calibration regressions and show that they can be interpreted as tests of correct forecast scale (Chetty et al., 2014a). We then discuss a test for within-year variation in test-score residuals (Chetty et al., 2011). Finally, we show how forecast aggregation can be used to test for across-teacher information in test-score residuals (Rothstein, 2017).

3.1 Forecast Calibration Regressions and Testing for Correct Forecast Scale

We can test whether a forecast is correctly calibrated to test-score residuals by estimating an OLS regression of test-score residuals on the forecast. These tests are useful when the researcher constructs forecasts using an information set that differs from the one used in the original model. For example, Chetty et al. (2014a) estimate a time-varying value-added model based on the full autocovariance matrix of test-score residuals. In their analysis, however, the forecasts are based on a leave-year-out information set, which is constructed from only part of this matrix. To demonstrate that their leave-year-out forecasts are correctly calibrated, they regress test-score residuals on these forecasts.[4]

To formalize this idea, suppose a value-added model produces forecasts, $\hat{A}_{it}^{(k)} = w^{(k)} M_{it}^{(k)}$, based on information set k. If the researcher instead wants to construct forecasts based on a different information set, k', then the forecast weight must be modified to account for the change in the information set. Denote this modified weight by $w^{(k|k')}$ and the candidate forecast by $\hat{A}_{it}^{(k|k')} = w^{(k|k')} M_{it}^{(k')}$. To test whether this candidate forecast is calibrated to the test-score residuals, we estimate

$$A_{it} = \alpha + \lambda^{(k')} \hat{A}_{it}^{(k|k')} + \zeta_{it}. \tag{3}$$

Since $\hat{A}_{it}^{(k|k')} = w^{(k|k')} M_{it}^{(k')}$,

$$\lambda^{(k')} = \frac{\text{Cov}\left(A_{it}, w^{(k|k')} M_{it}^{(k')}\right)}{\text{Var}\left(w^{(k|k')} M_{it}^{(k')}\right)} = \frac{w^{(k')}}{w^{(k|k')}}. \tag{4}$$

Following Mincer and Zarnowitz (1969), $\lambda^{(k')} = 1$ means that the candidate forecast is correctly calibrated to the test-score residuals. Equation (4) shows that this occurs exactly when the model-implied forecast weight, $w^{(k|k')}$, equals the correct forecast weight for the new information set, $w^{(k')}$. Thus, if the model-implied forecast weight differs from the correct weight, the candidate forecast will not be correctly calibrated to the outcome.

We illustrate this empirically using North Carolina elementary school data (North Carolina Department of Public Instruction, 2011), which are described in Appendix A and are similar to the data in Rothstein (2017). We estimate value-added models separately for reading and math using the Stata code provided by Chetty et al. (2014a). The program's standard output provides test-score residuals as well as leave-year-out forecasts based on

[4] See, for example, Table 3 in Chetty et al. (2014a).

Table 2: Testing Forecast Scale under Different Information Sets

	LYO	LOO	LYO†
	(1)	(2)	(3)
$\hat{\lambda}$	1.022	0.994	1.025
	(0.002)	(0.001)	(0.002)
Observations	5,150,990	7,093,236	7,093,236
R^2	0.050	0.090	0.036

† Students of one-year teachers are assigned a test-score forecast of zero.
Note: The table reports estimates from regressions of test-score residuals on predicted test-score residuals, under different information sets for the value-added model in Chetty et al. (2014a). Leave-one-out (LOO) forecasts include data from all students taught by the teacher except the score being forecast, while leave-year-out (LYO) forecasts exclude all students taught by the same teacher in that year. Standard errors are reported in parentheses.

the model, contained in the variable `tv`. To test whether the provided forecast has been correctly scaled for the leave-year-out information set, Column 1 in Table 2 reports the estimate of Eq. (3) using this forecast. The regression coefficient is 1.022, indicating that the model-implied forecast weight is close to the correct weight for the leave-year-out information set.[5]

In Column 2, we estimate Eq. (3) using leave-one-out forecasts constructed from the full autocovariance matrix, rather than the leave-year-out forecasts in Column 1, which use only the part of the matrix corresponding to other years. The sample size in this regression increases by roughly 40% because teachers observed for only one year are excluded from the leave-year-out forecast but can be included in the leave-one-out forecast. The R^2 also increases substantially, reflecting the inclusion of within-year variation in the forecast. For the leave-one-out forecast, the regression coefficient is 0.994, indicating a very close alignment with the correct weight. This is expected since this more inclusive forecast uses an information set closer to the one used to estimate the model.

Finally, Column 3 considers the treatment of teachers who are observed for only one year when constructing leave-year-out forecasts. This regression uses the leave-year-out forecasts but includes one-year teachers and assigns them a forecast of zero, the population mean of test-score residuals (Chetty et al., 2014a; Rothstein, 2017). If the average residual for these teachers differed from zero in a meaningful way, then assigning them a zero forecast

[5]Appendix Table A3 in Rothstein (2017) presents a very similar coefficient, 1.021.

should affect the calibration coefficient. The coefficient in Column 3 is nearly identical to the coefficient in Column 1, indicating that either treatment of one-year teachers produces similarly calibrated forecasts at the student level.

3.2 Testing for Within-Year Variation

Many value-added models pool within- and across-year variation but do not separately identify the sources of this variation, such as the value-added model discussed in Koedel et al. (2015). For these models, applying the test in Eq. (3) using the leave-year-out information set can provide insight into the presence of within-year variation in test-score residuals. In a pooled model, the forecast weight, $w^{(k)}$, is based on all T years of data. Given the leave-year-out information set, M_{it}^{LYO}, let $w^{(k|\text{LYO})}$ denote the corresponding model-implied forecast weight that adjusts only for the smaller number of years in the information set, without accounting for the fact that current-year information has been removed. Using the forecast $\hat{A}_{it}^{(k|\text{LYO})} = w^{(k|\text{LYO})} M_{it}^{\text{LYO}}$ in Eq. (3),

$$\lambda^{\text{LYO}} = \frac{\text{Cov}\left(A_{it}, w^{(k|\text{LYO})} M_{it}^{\text{LYO}}\right)}{\text{Var}\left(w^{(k|\text{LYO})} M_{it}^{\text{LYO}}\right)} = \frac{w^{\text{LYO}}}{w^{(k|\text{LYO})}}. \tag{5}$$

The numerator in Eq. (5) is the leave-year-out forecast weight, which is based only on across-year variation. The denominator is the sample-size-adjusted weight from the pooled model, which reflects both within- and across-year variation. In this model, the two weights are equal only when there is no transitory component of teacher value added and no classroom-level shocks. Thus, testing $\lambda^{\text{LYO}} = 1$ provides a direct test of no within-year variation in the test-score residuals. Rejection indicates that the sample-size-adjusted pooled weight differs from the leave-year-out weight, providing evidence of within-year variation in the test-score residuals.[6]

To see this interpretation concretely, consider a model that produces forecast weights for the leave-one-out information set M^{LOO} following Table 1:

$$w^{\text{LOO}} = \frac{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T}}{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T} + \frac{\sigma_\theta^2}{T} + \frac{\sigma_\varepsilon^2}{\tilde{n}T}}.$$

Because M^{LOO} includes current-year data, a share $1/T$ of the information set contains within-

[6]However, failing to reject $\lambda^{\text{LYO}} = 1$ does not provide direct evidence that within-year variation is absent. Since Eq. (5) is the ratio of two forecast weights, a precisely estimated coefficient of one means that within-year variation does not affect the forecast weight. This could occur because within-year variation is absent, or because the across-year signal is sufficiently large that within-year variation contributes little to the overall forecast.

year variation that is treated as signal. This variation provides predictive information about both the transitory teacher component and classroom shocks.

Adjusting this weight for the leave-year-out information set by accounting only for the reduction in the number of years from T to $T-1$, and not for the removal of within-year variation, gives the model-implied forecast weight

$$w^{(\text{LOO}|\text{LYO})} = \frac{\gamma_\mu + \frac{\sigma^2_\mu - \gamma_\mu}{T-1} + \frac{\sigma^2_\theta}{T-1}}{\gamma_\mu + \frac{\sigma^2_\mu - \gamma_\mu}{T-1} + \frac{\sigma^2_\theta}{T-1} + \frac{\sigma^2_\varepsilon}{\tilde{n}(T-1)}}.$$

This adjustment accounts for the smaller sample, but the numerator continues to treat a share $1/(T-1)$ of the information set as containing within-year variation, even though the omitted year is the current year and that variation has been removed.

The regression coefficient, λ^{LYO}, compares the correct forecast weight for the leave-year-out information set, w^{LYO}, in Table 1 with this model-implied weight, $w^{(\text{LOO}|\text{LYO})}$. This ratio simplifies to

$$\lambda^{\text{LYO}} = \frac{w^{\text{LYO}}}{w^{(\text{LOO}|\text{LYO})}} = \frac{\gamma_\mu}{\gamma_\mu + \frac{\sigma^2_\mu - \gamma_\mu}{T-1} + \frac{\sigma^2_\theta}{T-1}}.$$

Thus, in this case, testing $\lambda^{\text{LYO}} = 1$ is a test of the null hypothesis that $\gamma_\mu = \sigma^2_\mu + \sigma^2_\theta$.[7] Since $\gamma_\mu \leq \sigma^2_\mu$ and $\sigma^2_\theta \geq 0$, under these restrictions, this null hypothesis is equivalent to the joint hypothesis that there is no transitory component of teacher value added and no classroom-level shocks, $\sigma^2_\mu = \gamma_\mu$ and $\sigma^2_\theta = 0$. If there is either a transitory component of value added or classroom-level shocks, then $\lambda^{\text{LYO}} < 1$. In this case, rejecting the null indicates that the model-implied forecast weight, $w^{(\text{LOO}|\text{LYO})}$, places weight on within-year variation that is absent from the leave-year-out information set. As a result, the model-implied weight is greater than the correct forecast weight for the leave-year-out information set, pushing λ^{LYO} below one.

We implement this test using the North Carolina elementary school data, estimating the standard time-invariant value-added model described in Koedel et al. (2015), which we refer to as the pooled VAM. The pooled VAM estimates teacher value added from a teacher's average residual across all years taught, thereby pooling within- and across-year variation.[8] Since this model does not separately identify within- and across-year variation, applying the forecast regression in Eq. (3) to leave-year-out forecasts provides a test for within-year

[7] To simplify notation, we replace $\tilde{n}$ with n in the denominator of $w^{(\text{LOO}|\text{LYO})}$. This suppresses a small multiplicative factor, approximately equal to one, that reflects the marginal difference between the denominators of w^{LYO} and $w^{(\text{LOO}|\text{LYO})}$.

[8] The model in Koedel et al. (2015) cannot be directly mapped to the test-score components in Eq. (1). In their framework, test scores are modeled as $A_{it} = \tilde{\mu}_j + \tilde{\varepsilon}_{it}$. For a teacher observed for T years teaching n

Table 3: Forecast Regressions for the Pooled VAM under Different Information Sets

	LYO	LOO
	(1)	(2)
$\hat{\lambda}$	0.804	0.995
	(0.002)	(0.001)
Observations	5,150,990	7,093,236
R^2	0.044	0.073

Note: The table reports estimates from regressions of test-score residuals on predicted test-score residuals under different information sets for the time-invariant value-added model in Koedel et al. (2015). Leave-one-out (LOO) forecasts include data from all students taught by the teacher except the score being forecast, while leave-year-out (LYO) forecasts exclude all students taught by the same teacher in that year. Standard errors are reported in parentheses.

variation.

Column 1 in Table 3 shows the estimate of λ^{LYO} when the leave-year-out forecast is applied to the pooled VAM.[9] The regression coefficient, 0.804, is far below one, indicating that the leave-year-out forecast weight implied by the pooled VAM is substantially larger than the correct leave-year-out weight. In terms of the model in Eq. (1), this means that either teacher value added has a transitory component or classroom shocks are present, though the test cannot distinguish between the two. This result is similar to Chetty et al. (2011), the working paper version of Chetty et al. (2014a), which obtains a coefficient of 0.861 when applying a leave-year-out forecast to a model that pools within- and across-year variation.

It is also useful to view the results in Column 1 as a forecast calibration regression, which indicates that, in our data, leave-year-out forecasts from the pooled VAM are not correctly calibrated to test-score residuals and therefore should not be used in analyses that rely on

students each year, this implies leave-one-out and leave-year-out forecasts of

$$\hat{A}_{it}^{\text{LOO}} = \frac{\sigma_{\tilde{\mu}}^2}{\sigma_{\tilde{\mu}}^2 + \frac{\sigma_{\tilde{\varepsilon}}^2}{nT-1}} M_{j,it}^{\text{LOO}}, \qquad \hat{A}_{it}^{\text{LYO}} = \frac{\sigma_{\tilde{\mu}}^2}{\sigma_{\tilde{\mu}}^2 + \frac{\sigma_{\tilde{\varepsilon}}^2}{n(T-1)}} M_{j,it}^{\text{LYO}}.$$

Various approaches have been suggested for estimating the variance component $\sigma_{\tilde{\mu}}^2$. In this example, we follow the approach outlined in Koedel (2009).

[9] We estimate value-added models separately for reading and math and pool both subjects in these regressions.

forecasts excluding contemporaneous classroom data.[10] The coefficient below one reflects substantial within-year variation in the test-score residuals. When the pooled VAM is used to construct a leave-year-out forecast, the model-implied forecast weight remains too large because the model does not distinguish the within-year variation that has been removed from the across-year variation that remains. Forecasts from the pooled VAM should therefore be constructed using an information set closer to the one used to estimate the model, which includes both within- and across-year variation. For example, Column 2 uses leave-one-out forecasts in the regression, resulting in a coefficient of 0.995. This shows that the pooled VAM produces correctly calibrated forecasts for this more inclusive information set.

Finally, the estimate of λ^{LYO} in Column 1 should not be over-interpreted. While it indicates the presence of within-year variation and points to limitations in applying leave-year-out forecasts to the pooled VAM, this result by itself cannot determine which value-added model or forecast should be used more broadly. It does not imply that the pooled VAM should be rejected, nor does it imply that leave-year-out forecasts are preferred simply because classroom shocks may be present. Answering these broader questions requires shifting from forecast calibration to forecast choice, where the relevant trade-off depends on the policy application. This issue is the focus of Section 5.

3.3 Testing for Across-Teacher Information

Many forecast-based tests aggregate teacher-level forecasts to the school-grade or school level. For example, the teacher-switching test in Chetty et al. (2014a) aggregates teacher-level forecasts to the school-grade level. Rothstein (2017) highlights the potential role of across-teacher information when constructing school-level forecasts by averaging individual teacher-level forecasts, but argues that this issue is unlikely to be important in practice. We use our framework to show how averages of teacher-level forecasts can be used to directly test whether across-teacher information in value added is relevant for aggregate forecasts.

Define the average test-score residual in school s, grade g, and year t as the average of teacher-level averages:

$$\bar{A}_{sgt} = \frac{1}{J_{sgt}} \sum_{j \in \mathcal{J}_{sgt}} \bar{A}_{jt}, \tag{6}$$

where $\mathcal{J}_{sgt}$ is the set of teachers in school s, grade g, and year t, and J_{sgt} is the number of teachers in that set. Let $\hat{A}_{jt}^{(k)} = w^{(k)} M_{jt}^{(k)}$ denote the model's test-score forecast for teacher

[10] For example, the teacher-switching test in Chetty et al. (2014a), which relies on forecasts that exclude contemporaneous classroom data.

j in year t under information set k. The school-grade-year average of these forecasts is

$$Q_{sgt}^{(k)} = \frac{1}{J_{sgt}} \sum_{j \in \mathcal{J}_{sgt}} \hat{A}_{jt}^{(k)} = w^{(k)} \bar{M}_{sgt}^{(k)}, \tag{7}$$

where $\bar{M}_{sgt}^{(k)} = J_{sgt}^{-1} \sum_{j \in \mathcal{J}_{sgt}} M_{jt}^{(k)}$. Thus, averaging teacher-level forecasts is equivalent to applying the teacher-level forecast weight to the average information set.

We can assess whether this average forecast is correctly calibrated to the average residual by estimating

$$\bar{A}_{sgt} = \alpha + \lambda^{\text{agg}} Q_{sgt}^{(k)} + \zeta_{sgt}. \tag{8}$$

Let

$$\overline{w}^{(k)} = \frac{\text{Cov}\left(\bar{A}_{sgt}, \bar{M}_{sgt}^{(k)}\right)}{\text{Var}\left(\bar{M}_{sgt}^{(k)}\right)}$$

denote the coefficient for the best linear predictor of the average school-grade test-score residual given the average information set. Since $Q_{sgt}^{(k)} = w^{(k)} \bar{M}_{sgt}^{(k)}$, the coefficient in Eq. (8) is

$$\lambda^{\text{agg}} = \frac{\text{Cov}\left(\bar{A}_{sgt}, w^{(k)} \bar{M}_{sgt}^{(k)}\right)}{\text{Var}\left(w^{(k)} \bar{M}_{sgt}^{(k)}\right)} = \frac{\overline{w}^{(k)}}{w^{(k)}},$$

which shows that λ^{agg} is determined by the ratio between the best linear predictor of the average school-grade test-score residual given the average information set, $\overline{w}^{(k)}$, and the weight used to construct the teacher-level forecasts, $w^{(k)}$. Thus, $\lambda^{\text{agg}} = 1$ when these two weights are equal, indicating that the average forecast is correctly calibrated to the average test-score residual.

As shown in Eq. (2), the teacher-level weight, $w^{(k)}$, is the sum of the forecast weights for the individual test-score components. We can similarly decompose $\overline{w}^{(k)}$ into its component weights. If the test-score residuals follow Eq. (1), the average residual can be written as

$$\bar{A}_{sgt} = \bar{\mu}_{sgt} + \bar{\theta}_{sgt} + \bar{\varepsilon}_{sgt},$$

with each term denoting the school-grade average of the corresponding test-score component. Then

$$\overline{w}^{(k)} = \bar{w}_{\mu}^{(k)} + \bar{w}_{\theta}^{(k)} + \bar{w}_{\varepsilon}^{(k)},$$

where $\bar{w}_{\mu}^{(k)} = \text{Cov}(\bar{\mu}_{sgt}, \bar{M}_{sgt}^{(k)}) / \text{Var}(\bar{M}_{sgt}^{(k)})$, with $\bar{w}_{\theta}^{(k)}$ and $\bar{w}_{\varepsilon}^{(k)}$ defined analogously. Expressing

λ^{agg} in terms of these component weights gives

$$\lambda^{\text{agg}} = \frac{\bar{w}_{\mu}^{(k)} + \bar{w}_{\theta}^{(k)} + \bar{w}_{\varepsilon}^{(k)}}{w_{\mu}^{(k)} + w_{\theta}^{(k)} + w_{\varepsilon}^{(k)}}. \tag{9}$$

Estimating Eq. (8) and interpreting its coefficient can be challenging because Eq. (9) shows that λ^{agg} depends on multiple component weights. Testing the null hypothesis that $\lambda^{\text{agg}} = 1$ is a test of whether the correct total forecast weight for the aggregate residual equals the teacher-level total weight, $\overline{w}^{(k)} = w^{(k)}$. Because the aggregate component weights, $\bar{w}_{\mu}^{(k)}$, $\bar{w}_{\theta}^{(k)}$, and $\bar{w}_{\varepsilon}^{(k)}$, are based on an information set that pools information across teachers, across-teacher information in the test-score components will generally cause them to differ from the corresponding teacher-level component weights. In this case, $\lambda^{\text{agg}} \neq 1$ provides evidence of across-teacher information in the data but does not identify in which component. At the same time, because differences across components can offset one another, $\lambda^{\text{agg}} = 1$ does not imply the absence of across-teacher information. Thus, a deeper interpretation of λ^{agg} requires additional structure that links the coefficient to the underlying test-score components.

The interpretation of λ^{agg} can be made more straightforward by using a leave-year-out information set, which eliminates some of the component weights in Eq. (9). Because the leave-year-out information set contains no signal about contemporaneous classroom shocks or student-level shocks, the aggregate component weights, $\bar{w}_{\theta}^{\text{LYO}}$ and $\bar{w}_{\varepsilon}^{\text{LYO}}$, and the teacher-level component weights, w_{θ}^{LYO} and $w_{\varepsilon}^{\text{LYO}}$, are all equal to zero.

Thus, in the leave-year-out case, $\lambda^{\text{agg}} = 1$ when $w_{\mu}^{\text{LYO}} = \bar{w}_{\mu}^{\text{LYO}}$. Since the school-grade average information set pools teachers within schools, $\bar{M}_{sgt}^{\text{LYO}}$ may contain signal about $\bar{\mu}_{sgt}$ not only from each teacher's own past residuals, but also from the residuals of other teachers in the same school and grade. If teacher value added is correlated within schools, this across-teacher information enters the forecast weight for the aggregate residual, $\bar{w}_{\mu}^{\text{LYO}}$, causing it to differ from the teacher-level value-added weight, w_{μ}^{LYO}, which is based only on within-teacher variation. As a result, a coefficient different from one in the leave-year-out aggregate forecast regression provides direct evidence of across-teacher information in value added.

To illustrate this using the framework in Section 2, let $\delta_{\mu} = \text{Cov}(\mu_{jt}, \mu_{j't'})$ for $j \neq j'$ denote the covariance in value added across distinct teachers within the same school, assumed for simplicity to be constant across the relevant years. If $\bar{M}_{sgt}^{\text{LYO}}$ is the average of the leave-year-out information set used in Table 1, then the coefficient of the best linear predictor of $\bar{A}_{sgt}$ given $\bar{M}_{sgt}^{\text{LYO}}$ is

$$\overline{w}^{\text{LYO}} = \frac{\gamma_{\mu} + (J_{sgt} - 1)\delta_{\mu}}{\gamma_{\mu} + \frac{\sigma_{\mu}^2 - \gamma_{\mu}}{T-1} + \frac{\sigma_{\theta}^2}{T-1} + \frac{\sigma_{\varepsilon}^2}{n(T-1)} + (J_{sgt} - 1)\delta_{\mu}}. \tag{10}$$

Comparing this expression to the teacher-level weight, w^{LYO}, in Table 1 shows that $\overline{w}^{\text{LYO}} = w^{\text{LYO}}$ only when $\delta_\mu = 0$. Thus, in this case, estimating Eq. (8) and testing whether $\lambda^{\text{agg}} = 1$ is a test of whether $\delta_\mu = 0$. If teacher value added is positively correlated within schools, so that $\delta_\mu > 0$, the aggregate-level weight is larger than the teacher-level weight, $\overline{w}^{\text{LYO}} > w^{\text{LYO}}$, and $\lambda^{\text{agg}} > 1$.[11]

Next, we estimate the aggregate forecast regression in Eq. (8) to assess the empirical relevance of across-teacher information in teacher value added using the North Carolina elementary school data. For this analysis, we use the value-added model and corresponding leave-year-out forecasts used in Table 2, which are based on the Stata code provided by Chetty et al. (2014a). We average student-level test-score residuals to the school-grade-year level, along with the corresponding leave-year-out teacher-level forecasts based on the variable `tv`. To analyze how the treatment of teachers observed for only one year affects aggregation, we conduct the analysis two ways: first excluding these teachers, and second including them and assigning them a teacher-level forecast of zero. As shown in Table 2, either treatment produces correctly calibrated teacher-level forecasts, so differences between the two aggregate regressions can be attributed to aggregation.

Column 1 in Table 4 reports the aggregate regression coefficient when teachers with only a single year of data are removed from the analysis. Since this aggregation is based on leave-year-out forecasts, the regression isolates across-teacher information in value added at the school-grade level. The coefficient estimate, 1.188, differs from one, which provides evidence of across-teacher information in value added. Because the coefficient is greater than one, Eq. (10) implies that teacher value added is positively correlated within schools, so that the correct aggregate forecast weight is larger than the teacher-level weight used in the averaged forecast.[12]

From a calibration perspective, Column 1 also demonstrates that, due to the presence of across-teacher information in value added, the average of the teacher-level forecasts is incorrectly calibrated to the average school-grade test-score residual. The estimate of 1.188 implies that the correct aggregate forecast weight is 18.8% larger than the teacher-level weight used to construct the averaged forecast. Thus, in this case, averaged teacher-level forecasts should not be used for analyses that require correctly calibrated school-grade-level

[11] As in the within-year test discussed in Footnote 6, failing to reject $\lambda^{\text{agg}} = 1$ does not provide direct evidence that across-teacher information is absent. Since Eq. (9) is the ratio of two forecast weights, a precisely estimated coefficient of one means that across-teacher information does not affect the aggregate forecast weight. This could occur because across-teacher information is absent, or because the within-teacher signal is sufficiently large that across-teacher information contributes little to the overall forecast.

[12] If test-score residuals instead follow Eq. (11), then $\lambda^{\text{agg}} \neq 1$ reflects across-teacher information in the combined persistent components of test-score residuals, which could include both teacher value added and persistent non-teacher inputs. The test does not distinguish between the two sources.

Table 4: Forecast Aggregation and Across-Teacher Information

	Single-Year Teachers	
	Remove	Assign Zero
	(1)	(2)
$\hat{\lambda}$	1.188	1.337
	(0.007)	(0.009)
Observations	46,318	51,127
R^2	0.394	0.299

Note: The table reports estimates from aggregate forecast regressions using leave-year-out value-added forecasts estimated separately for math and reading and pooled across subjects. Both test-score residuals and forecasts are averaged to the school-grade-year level. Column 1 excludes teachers observed for only one year. Column 2 includes these teachers and assigns them a teacher-level forecast of zero. Standard errors are reported in parentheses.

forecasts. Column 2 shows that the averaged teacher-level forecast is even more poorly calibrated when one-year teachers are included and assigned a forecast of zero. In this case, the coefficient rises to 1.337, implying a larger gap between the teacher-level forecast weight, w^{LYO}, and the correct aggregate weight, $\overline{w}^{\text{LYO}}$. This occurs because teachers with little data have limited within-teacher signal, making the across-teacher signal introduced by aggregation more important.

These results show that while Rothstein (2017) argues that within-school correlations in teacher value added are likely too small to meaningfully affect forecast aggregation, the effect of across-teacher information on forecast aggregation depends on how much signal this information adds relative to the available within-teacher signal. This cannot be assessed from the size of the within-school correlation alone. Even a small within-school correlation in teacher value added can have a meaningful effect on forecast aggregation when the number of observations per teacher is small. In the North Carolina data, depending on the treatment of one-year teachers, the correct aggregate forecast weight is 18.8% to 33.7% larger than the weight used in the averaged teacher-level forecasts. Thus, estimates of Eq. (8) provide a more direct way to assess the empirical importance of across-teacher information for forecast aggregation.

4 The Chetty et al. (2014a) Teacher-Switching Test

A central concern in teacher value-added models is whether teacher-level forecasts are biased by persistent non-teacher inputs. In the absence of randomized teacher assignment, as in Kane and Staiger (2008), or auxiliary family-background information, as in Chetty et al. (2014a), forecast bias is difficult to assess directly. The most notable approach for testing bias is the teacher-switching test developed by Chetty et al. (2014a). The test is influential because Chetty et al. (2014a) argue that it provides a quasi-experimental test of forecast bias that can be implemented with administrative data in lieu of randomized assignment. Although there is debate about the interpretation of the test (Rothstein, 2017; Chetty et al., 2017), it has become a benchmark in the empirical and methodological literature on value added (Bates et al., 2025; Delgado, 2025; Gilraine and McCarthy, 2024; Gilraine and Pope, 2021; Backes and Hansen, 2018; Bacher-Hicks et al., 2014; Chetty et al., 2014b).

Building on the previous section, we apply our framework to analyze the teacher-switching test of Chetty et al. (2014a), showing what the switching coefficient recovers and when it can be interpreted as a test of teacher-level forecast bias. To allow for persistent non-teacher inputs, we introduce an additional component into the test-score residual, similar to Rothstein (2017):

$$A_{it} = \mu_{jt} + \pi_{jt} + \theta_{jt} + \varepsilon_{it}. \tag{11}$$

As before, μ_{jt} represents teacher value added. The new term, π_{jt}, captures unobserved classroom or student inputs that affect achievement but are not attributable to the teacher. Unlike the classroom shock θ_{jt}, which is transitory, π_{jt} may persist over time if teachers are repeatedly assigned students with similar unobserved characteristics, so that $\text{Cov}(\pi_{jt}, \pi_{jt'}) \neq 0$.

To see how π_{jt} creates potential forecast bias, consider the forecast weight for the leave-year-out information set in Table 1 when test-score residuals follow Eq. (11). Let $\gamma_\pi = \text{Cov}(\pi_{jt}, \pi_{jt'})$ for $t \neq t'$ denote the covariance of non-teacher inputs across time. Since student assignment may be correlated with teacher value added, let $\tau = \text{Cov}(\mu_{jt}, \pi_{jt'})$ denote the covariance between teacher value added and persistent non-teacher inputs, which we assume is constant across years to simplify notation. Then the forecast weight for the leave-year-out information set, which relies only on across-year variation, is

$$w^{\text{LYO}} = w_\mu^{\text{LYO}} + w_\pi^{\text{LYO}} = \frac{\gamma_\mu + \tau}{\text{Var}\left(M_{jt}^{\text{LYO}}\right)} + \frac{\gamma_\pi + \tau}{\text{Var}\left(M_{jt}^{\text{LYO}}\right)}. \tag{12}$$

This equation shows that persistent non-teacher inputs affect the forecast in two ways. First,

they enter as a forecast component through w_π^{LYO}, meaning that π_{jt} is part of the prediction target. Second, when $\tau \neq 0$, persistent non-teacher inputs are informative about teacher value added and therefore also enter as signal about μ_{jt}. This is why the covariance term τ appears in w_μ^{LYO}.

As discussed in Section 2, a test-score forecast is equivalent to a value-added forecast when value added is the only component in the forecast, so that $w^{(k)} = w_\mu^{(k)}$. This condition corresponds to what Chetty et al. (2014a) and Rothstein (2017) refer to as forecast unbiasedness. Thus, forecast unbiasedness requires the leave-year-out test-score forecast to place no weight on persistent non-teacher inputs, so $w_\pi^{\text{LYO}} = 0$, which occurs when $\gamma_\pi + \tau = 0$. Forecast unbiasedness only requires that π_{jt} not be part of the prediction target, but it does not preclude non-teacher inputs from serving as a source of information about teacher value added.

Chetty et al. (2014a) define a forecast as teacher-level unbiased if the forecast converges to the teacher's true value added as the number of classes increases. For a forecast to be teacher-level unbiased, the test-score variation used to identify μ_{jt} must come entirely from variation in μ_{jt} and no other source. Thus, while forecast unbiasedness requires $\gamma_\pi + \tau = 0$, teacher-level unbiasedness additionally requires $\tau = 0$. Together, these conditions imply $\gamma_\pi = 0$, so teacher-level unbiasedness occurs only when there is no persistent non-teacher component in test scores. Since π_{jt} is defined as the persistent non-teacher component of test scores, any non-teacher variation that does not persist over time is captured by θ_{jt} rather than π_{jt}. Thus, in this framework, teacher-level unbiasedness equivalently requires the variance of the persistent non-teacher component to equal zero, $\sigma_\pi^2 = 0$, which corresponds to the definition in Rothstein (2017).

The test proposed by Chetty et al. (2014a) for forecast unbiasedness uses teachers entering and exiting school-grades as a quasi-experiment. The test is based on forecasts of year-over-year changes in average school-grade test-score residuals. This change is given by

$$\Delta \bar{A}_{sgt} = \bar{A}_{sgt} - \bar{A}_{sg,t-1},$$

where $\bar{A}_{sgt}$ is defined in Eq. (6). Using the average school-grade-year teacher-level forecast, $Q_{sgt}^{(k)}$, from Eq. (7), the model-implied forecast of this change is constructed as the year-over-year change in the average teacher-level forecast,

$$\Delta Q_{sgt}^{(k)} = Q_{sgt}^{(k)} - Q_{sg,t-1}^{(k)} = w^{(k)}\, \Delta \bar{M}_{sgt}^{(k)},$$

where $\Delta \bar{M}_{sgt}^{(k)} = \bar{M}_{sgt}^{(k)} - \bar{M}_{sg,t-1}^{(k)}$. This equation shows that differencing the averaged teacher-level forecasts is equivalent to applying the teacher-level forecast weight to the year-over-year

change in the averaged information set.

The test regresses the actual change in the average test-score residuals on the model-implied change in forecasted test scores:

$$\Delta \bar{A}_{sgt} = \Delta \alpha_t + \lambda^{\text{CFR}} \Delta Q^{(k)}_{sgt} + \Delta \chi_{sgt}. \tag{13}$$

Chetty et al. (2014a) argue that λ^{CFR} measures teacher-level forecast bias in the model and use $\lambda^{\text{CFR}} = 1$ to indicate unbiased teacher-level forecasts.

To see how λ^{CFR} relates to forecast unbiasedness, let

$$b^{(k)} = \frac{\text{Cov}\left(\Delta \bar{A}_{sgt}, \Delta \bar{M}^{(k)}_{sgt}\right)}{\text{Var}\left(\Delta \bar{M}^{(k)}_{sgt}\right)}$$

be the coefficient of the best linear predictor of the change in the average school-grade test-score residual given the change in the averaged information set. Since $\Delta Q^{(k)}_{sgt} = w^{(k)} \Delta \bar{M}^{(k)}_{sgt}$, the coefficient in Eq. (13) is

$$\lambda^{\text{CFR}} = \frac{\text{Cov}\left(\Delta \bar{A}_{sgt}, w^{(k)} \Delta \bar{M}^{(k)}_{sgt}\right)}{\text{Var}\left(w^{(k)} \Delta \bar{M}^{(k)}_{sgt}\right)} = \frac{b^{(k)}}{w^{(k)}}.$$

This expression shows that λ^{CFR} is the ratio of the best linear predictor coefficient for changes in school-grade test-score residuals to the model-implied forecast weight used to construct the teacher-level forecasts. The numerator of this ratio can be related to the underlying test-score components by decomposing the change in average test-score residuals using the components in Eq. (11):

$$\Delta \bar{A}_{sgt} = \Delta \bar{\mu}_{sgt} + \Delta \bar{\pi}_{sgt} + \Delta \bar{\theta}_{sgt} + \Delta \bar{\varepsilon}_{sgt},$$

where each term denotes the year-over-year change in the school-grade average of the corresponding component. Then, similar to the decomposition of the forecast weight in Eq. (2), $b^{(k)}$ can be decomposed into the sum of the best linear predictor coefficients for each component:

$$b^{(k)} = b^{(k)}_{\mu} + b^{(k)}_{\pi} + b^{(k)}_{\theta} + b^{(k)}_{\varepsilon},$$

where $b^{(k)}_{\mu} = \text{Cov}(\Delta \bar{\mu}_{sgt}, \Delta \bar{M}^{(k)}_{sgt}) / \text{Var}(\Delta \bar{M}^{(k)}_{sgt})$, with $b^{(k)}_{\pi}$, $b^{(k)}_{\theta}$, and $b^{(k)}_{\varepsilon}$ defined analogously.

The teacher-switching test uses a leave-two-year-out information set, M^{L2YO}_{jt}. Like the leave-year-out information set, it contains no signal about contemporaneous shocks, so

$w_\theta^{\text{L2YO}} = 0$ and $w_\varepsilon^{\text{L2YO}} = 0$ in the teacher-level forecasts. Because two years are removed, it also contains no signal about changes in the average contemporaneous components, so

$$\text{Cov}(\Delta\bar{\theta}_{sgt}, \Delta\bar{M}_{sgt}^{\text{L2YO}}) = 0 \quad \text{and} \quad \text{Cov}(\Delta\bar{\varepsilon}_{sgt}, \Delta\bar{M}_{sgt}^{\text{L2YO}}) = 0.$$

Thus, $b_\theta^{\text{L2YO}} = 0$ and $b_\varepsilon^{\text{L2YO}} = 0$ in the coefficient for the best linear predictor. In the case of the leave-two-year-out information set, the coefficient for the teacher-switching test therefore simplifies to

$$\lambda^{\text{CFR}} = \frac{b_\mu^{\text{L2YO}} + b_\pi^{\text{L2YO}}}{w_\mu^{\text{L2YO}} + w_\pi^{\text{L2YO}}}. \tag{14}$$

Next, we use the framework in Section 2 to illustrate the determinants of λ^{CFR} in Eq. (14). To simplify the notation, assume that in a given school-grade there is only one teacher. In period $t-1$, that teacher is teacher 1, who is replaced in that school-grade in period t by teacher 2. Both teachers are observed in period $t+1$, from which the leave-two-year-out forecast can be constructed. In this case

$$\Delta\bar{A}_{sgt} = (\mu_{2t} - \mu_{1,t-1}) + (\pi_{2t} - \pi_{1,t-1}) + (\theta_{2t} - \theta_{1,t-1}) + (\bar{\varepsilon}_{2t} - \bar{\varepsilon}_{1,t-1}),$$

and

$$\Delta\bar{M}_{sgt}^{\text{L2YO}} = (\mu_{2,t+1} - \mu_{1,t+1}) + (\pi_{2,t+1} - \pi_{1,t+1}) + (\theta_{2,t+1} - \theta_{1,t+1}) + (\bar{\varepsilon}_{2,t+1} - \bar{\varepsilon}_{1,t+1}).$$

In addition to the within-teacher covariances defined above, γ_π and τ, define the across-teacher covariances for switching teacher pairs, $j \neq j'$, as

$$\delta_\mu^{\text{sw}} = \text{Cov}(\mu_{jt}, \mu_{j't'}), \qquad \delta_\pi^{\text{sw}} = \text{Cov}(\pi_{jt}, \pi_{j't'}), \qquad \psi^{\text{sw}} = \text{Cov}(\mu_{jt}, \pi_{j't'}).^{13}$$

The parameter δ_μ^{sw} captures the covariance in value added between switching teachers, δ_π^{sw} captures the covariance in persistent non-teacher inputs between switching teachers, and ψ^{sw} captures the covariance between one switching teacher's value added and the persistent non-teacher component assigned to the other switching teacher. Then the best linear predictor

[13] Rothstein (2017) (footnote 14) notes that aggregation of teacher-level forecasts requires independence not only across teachers within schools, but also between outgoing and incoming teachers. We make the implications of relaxing this restriction explicit for the teacher-switching coefficient and allow for dependence between switching teachers in both teacher value added and persistent non-teacher inputs, as well as across these components.

coefficient for changes in school-grade residuals is given by

$$b^{\text{L2YO}} = \frac{\gamma_\mu + \gamma_\pi + 2\tau - 2\psi^{\text{sw}} - \delta_\mu^{\text{sw}} - \delta_\pi^{\text{sw}}}{\sigma_\mu^2 + \sigma_\pi^2 + \sigma_\theta^2 + \frac{\sigma_\varepsilon^2}{n} + 2\tau - 2\psi^{\text{sw}} - \delta_\mu^{\text{sw}} - \delta_\pi^{\text{sw}}},$$

and the teacher-level forecast weight is

$$w^{\text{L2YO}} = \frac{\gamma_\mu + \gamma_\pi + 2\tau}{\sigma_\mu^2 + \sigma_\pi^2 + \sigma_\theta^2 + \frac{\sigma_\varepsilon^2}{n} + 2\tau}.$$

These expressions help clarify what can be tested in the teacher-switching regression. As stated above, the teacher-level forecasts are unbiased if $w_\pi^{\text{L2YO}} = 0$, which follows if $\gamma_\pi + \tau = 0$. A direct test of forecast unbiasedness would therefore test

$$H_0^A : \gamma_\pi + \tau = 0.$$

The teacher-switching regression alone cannot test H_0^A, because this hypothesis does not map into a unique value of λ^{CFR}. The null H_0^A restricts the sum $\gamma_\pi + \tau$, but the switching coefficient depends on γ_π and τ separately. For example, substituting $\gamma_\pi = -\tau$ into b^{L2YO} and w^{L2YO} still leaves τ in both expressions. Thus, the value of λ^{CFR} is not pinned down by forecast unbiasedness alone.

Using λ^{CFR} to test the more restrictive case of teacher-level unbiasedness corresponds to the null hypothesis

$$H_0^B : \sigma_\pi^2 = 0,$$

which states that there are no persistent non-teacher inputs. Under H_0^B, the covariances γ_π, τ, δ_π^{sw}, and ψ^{sw} are all zero, so H_0^B is a special case of H_0^A. Substituting these restrictions into b^{L2YO} and w^{L2YO} gives

$$\lambda^{\text{CFR}} = \frac{\gamma_\mu - \delta_\mu^{\text{sw}}}{\sigma_\mu^2 + \sigma_\theta^2 + \frac{\sigma_\varepsilon^2}{n} - \delta_\mu^{\text{sw}}} \frac{\sigma_\mu^2 + \sigma_\theta^2 + \frac{\sigma_\varepsilon^2}{n}}{\gamma_\mu}. \tag{15}$$

This expression shows that the value of λ^{CFR} implied by H_0^B depends on δ_μ^{sw}, which captures the relationship between the value added of the switching teachers. If entrants tend to replace similar teachers, then $\delta_\mu^{\text{sw}} > 0$ and the switching coefficient is less than one. If lower-value-added teachers are more likely to be dismissed and replaced by entrants drawn randomly from the population, then $\delta_\mu^{\text{sw}} < 0$ and the switching coefficient is greater than one. Unlike H_0^A, which does not map into a unique value of λ^{CFR}, H_0^B maps into a value of λ^{CFR} only conditional on δ_μ^{sw}. Therefore, using λ^{CFR} to test teacher-level unbiasedness

is difficult because the appropriate null value depends on the dataset-specific value of δ_μ^{sw}, which may be unknown.

If we use $\lambda^{\text{CFR}} = 1$ as the null value in the teacher-switching regression, as suggested by Chetty et al. (2014a), then Eq. (15) shows that this coincides with the correct value under H_0^B only if $\delta_\mu^{\text{sw}} = 0$. Since this condition may not hold in general, testing $\lambda^{\text{CFR}} = 1$ should be interpreted as a joint test of teacher-level unbiasedness and no covariance in the value added between entrants and the teachers they replace:

$$H_0^C : \sigma_\pi^2 = 0 \quad \text{and} \quad \delta_\mu^{\text{sw}} = 0.$$

There are two challenges with using $\lambda^{\text{CFR}} = 1$ as a test of forecast unbiasedness. The first is that this is a joint test, so the null may be rejected even when there are no persistent non-teacher inputs if $\delta_\mu^{\text{sw}} \neq 0$. The second is that $\lambda^{\text{CFR}} = 1$ is not sufficient for teacher-level forecast unbiasedness. Due to offsetting effects, many combinations of the covariance terms can produce $\lambda^{\text{CFR}} = 1$ without implying forecast unbiasedness. For example, in the simplified two-teacher case above, the least restrictive null corresponding to $\lambda^{\text{CFR}} = 1$ is

$$H_0^D : 2\psi^{\text{sw}} + \delta_\mu^{\text{sw}} + \delta_\pi^{\text{sw}} = 0.$$

This hypothesis places no restriction on $\gamma_\pi + \tau$, so $\lambda^{\text{CFR}} = 1$ can arise even when the teacher-level forecasts are biased. Thus, $\lambda^{\text{CFR}} = 1$ is not only insufficient for teacher-level unbiasedness, as argued by Rothstein (2017); it is also insufficient for forecast unbiasedness.

This analysis shows that the teacher-switching coefficient cannot be interpreted mechanically as a test of teacher-level forecast unbiasedness. A coefficient different from one does not necessarily imply biased teacher-level forecasts, because rejection may be driven by across-teacher information in value added among switching teachers. Similarly, a coefficient equal to one does not establish forecast unbiasedness, because offsetting covariance terms can produce $\lambda^{\text{CFR}} = 1$ even when the teacher-level forecast loads on persistent non-teacher inputs. Thus, without additional restrictions or assumptions, such as random assignment of teachers across schools to rule out across-teacher information, the test is best interpreted as a joint test whose conclusions cannot be decisively attributed to teacher-level forecast bias.

4.1 Monte Carlo Illustration

For the two-teacher case above, Eq. (15) shows that the correct null value for λ^{CFR} depends on the across-teacher information in value added for the switching teachers. If teachers are sorted across schools so that the entrant's value added is positively correlated with the value

added of the teacher being replaced, then $\delta_\mu > 0$ and Eq. (15) suggests $\lambda^{\text{CFR}} < 1$, even when value-added forecasts are unbiased for teacher value added.

To demonstrate that the same pattern arises outside the two-teacher case, we conduct a Monte Carlo experiment with multiple teachers per school and variation across schools in the number of switching teachers. The exercise is constructed so that teacher forecasts are unbiased for teacher value added, but teacher sorting causes the coefficient in the teacher-switching regression to differ from one. We simulate 5,000 schools with four teachers assigned to each school in each year, giving a total of 20,000 teachers observed over a 20-year period. For each teacher, we draw value added, $\mu_j \sim N(0, 0.05)$, and generate teacher-year average residuals according to

$$\bar{A}_{jt} = \mu_j + u_{jt}, \qquad u_{jt} \sim N\left(0, \frac{0.25}{20}\right),$$

for $t \in \{1, \ldots, 20\}$.

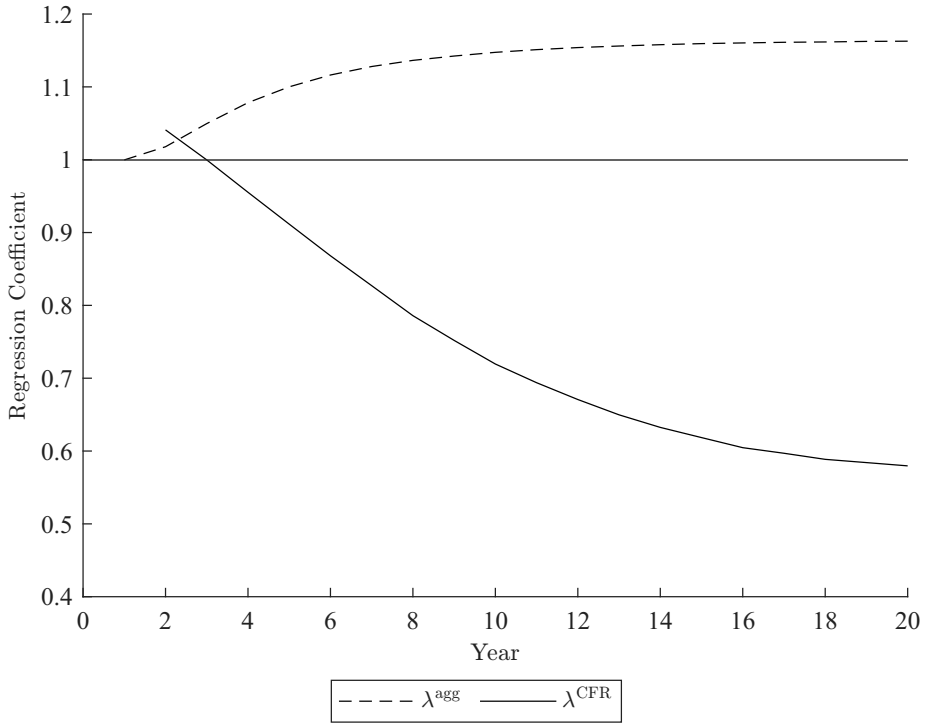
In the first year, teachers are assigned to schools at random. In each subsequent year, 20% of teachers are randomly selected to switch schools. We rank switching teachers by their prior-year average residual, $\bar{A}_{j,t-1}$, and rank schools by their index, $s \in \{1, 2, \ldots, 5000\}$. Lower-ranked switching teachers are assigned to openings in lower-indexed schools, while higher-ranked switching teachers are assigned to openings in higher-indexed schools. Repeating this process through $T = 20$ generates assortative matching across schools over time.

The simulation produces 20 years of teacher-switching data, with assignments initially close to random and becoming more sorted over time. For each year, we aggregate test-score residuals and teacher-level value-added forecasts to the school level. Using the aggregated school-level data, we first estimate λ^{agg} in Eq. (8) from Section 3 to measure the extent to which sorting creates across-teacher information in value added that affects the aggregate forecast.[14] We then take year-over-year differences in school-level test-score residuals and average teacher forecasts and estimate the teacher-switching coefficient λ^{CFR} in Eq. (13). Estimating these coefficients separately by year, rather than pooling all years, shows how aggregation and the switching test change as assortative matching accumulates over time.

Figure 1 reports the average coefficient from the aggregation regression, λ^{agg}, and the

[14]To isolate the effect of sorting on the coefficient estimates, we use a separate one-year average residual to construct each teacher's forecast rather than estimating value added using a leave-two-year-out procedure. This residual is not used elsewhere in the simulated panel and provides a single forecast that is held fixed for each teacher across all years. This construction ensures that the forecasts are unbiased for teacher value added, so changes in the coefficients are driven by the school assignment process rather than by forecast estimation.

Figure 1: Switching Test Coefficient under Assortative Teacher Sorting



Note: The figure reports average coefficients from 1,000 Monte Carlo simulations. The dashed line reports the coefficient from the aggregation regression, λ^{agg}, in Eq. (8). The solid line reports the teacher-switching coefficient, λ^{CFR}, in Eq. (13). Both coefficients are estimated separately by year in each simulation and then averaged across simulations.

average teacher-switching coefficient, λ^{CFR}, by year across 1,000 simulations. In the early periods, both coefficients are close to one because teacher assignments are still close to random. Over time, as sorting accumulates across schools, λ^{agg} rises above one. This is consistent with Eq. (10), since school-level forecasts increasingly contain across-teacher information that is not captured by the teacher-level forecast weight. At the same time, λ^{CFR} falls below one, consistent with Eq. (15), as entrants increasingly replace teachers with similar value added. By the end of the simulation, the coefficient from the aggregation regression is above 1.15, while the teacher-switching coefficient is below 0.6. Both coefficients move away from one even though the underlying value-added forecasts are unbiased for teacher value added. This shows that across-teacher sorting can change the correct null value for the teacher-switching coefficient, so deviations of λ^{CFR} from one cannot be interpreted mechanically as evidence of teacher-level forecast bias.

5 Forecast Choice in Education Policy

We now use our framework to examine how forecast choice affects education policy analysis. If test-score residuals follow Eq. (1), then researchers face a choice between forecasts that rely only on across-year test-score variation and forecasts that also include current-year data. As discussed in Section 2, leave-year-out forecasts can be appealing because they exclude contemporaneous classroom data and therefore avoid loading on contemporaneous classroom shocks. However, for certain applications there may be some appeal to using forecasts that also include current-year data. These forecasts contain more signal about teacher value added than information sets based only on across-year variation. This occurs for two reasons. First, they include an additional year of data, which provides more information about the persistent component of value added. Second, they incorporate the transitory component of value added as part of the forecast signal. In contrast, the leave-year-out forecast uses less data about the persistent component of value added and treats the transitory component of value added as noise.

This creates a central trade-off in forecast construction. Forecasts that isolate the teacher component more aggressively do so by discarding variation that is also informative about teacher value added. As a result, leave-year-out forecasts are cleaner with respect to contemporaneous classroom shocks, but they use less teacher signal. More inclusive forecasts retain more teacher signal, but may also load on shocks that are not attributable to the teacher. The forecast weights in Table 1 shows that this trade-off is especially pronounced in short panels. When there are fewer years of data, excluding a single year is more costly for the leave-year-out forecast. At the same time, classroom shocks may have a larger influence on

full-sample and leave-one-out forecasts because they are averaged over fewer years.

In this section, we compare how this trade-off affects two common applications of value-added models discussed in Hanushek and Rivkin (2012). First, we consider teacher ranking, where forecasts may be used for personnel decisions and accountability policies. Second, we consider estimates of the effects of teacher quality on long-term student outcomes.

5.1 Ranking Teachers

Teacher value-added estimates are widely used to evaluate teachers and inform workforce policies, many of which depend on teachers' relative positions in the value-added distribution (Hanushek, 2009). Ideally, teachers would be ranked based on their true value added, μ_{jt}, but because μ is unobserved, rankings must rely on forecasts. The trade-off developed above makes it unclear which forecasting approach should produce rankings most closely aligned with μ. Leave-year-out forecasts are orthogonal to contemporaneous classroom shocks, while more inclusive forecasts contain more value-added signal.

To evaluate which forecasting approach produces more accurate rankings, we derive the R^2 between each forecast and latent teacher value added. Given a forecast $\hat{A}^{(k)}$ constructed from information set $M^{(k)}$, this measure is

$$R^2_{\mu|\hat{A}^{(k)}} = \frac{\left(w_\mu^{(k)}\right)^2 \text{Var}\left(M^{(k)}\right)}{\sigma_\mu^2}.$$

This is not a measure of fit to observed test scores. It measures how closely the forecast aligns with latent teacher value added.[15] Since σ_μ^2 is common across forecasting approaches, comparisons across methods depend only on the numerator.

Substituting the component weights from Table 1, the leave-year-out forecast has

$$R^2_{\mu|\hat{A}^{\text{LYO}}} = w_\mu^{\text{LYO}} \cdot \frac{\gamma_\mu}{\sigma_\mu^2},$$

while the full-sample forecast has

$$R^2_{\mu|\hat{A}^{\text{FS}}} = w_\mu^{\text{FS}} \cdot \frac{\gamma_\mu + \frac{\sigma_\mu^2 - \gamma_\mu}{T}}{\sigma_\mu^2}.$$

Comparing the R^2 shows that the full-sample forecast is more closely aligned with μ_{jt}. First,

[15] Assessing the trade-off between leave-year-out and more inclusive forecasts is relevant only for in-sample forecasts. For out-of-sample forecasts, there is no contemporaneous classroom to leave out, so the natural approach is to use all available data.

$w_\mu^{\text{FS}} > w_\mu^{\text{LYO}}$, since the full-sample component weight has both a larger numerator and a smaller denominator.[16] Second, the full-sample forecast incorporates both the persistent component of teacher value added, γ_μ, and the transitory component, $(\sigma_\mu^2 - \gamma_\mu)/T$, whereas the leave-year-out forecast loads only on the persistent component, γ_μ.[17] Thus,

$$R^2_{\mu|\hat{A}^{\text{FS}}} > R^2_{\mu|\hat{A}^{\text{LYO}}}. \tag{16}$$

To illustrate the practical implications of this result, we conduct Monte Carlo exercises in which the goal is to identify teachers in the bottom 5% of the value-added distribution (Hanushek, 2009). Because μ_{jt} is unobserved, teachers are ranked using either a full-sample forecast, $\hat{A}^{\text{FS}}$, or a leave-year-out forecast, $\hat{A}^{\text{LYO}}$. To compare the performance of the two forecasting approaches, we evaluate the misclassification rate of each method, defined as the probability that a teacher is classified in the bottom 5% based on the forecast but does not belong to the bottom 5% of the true value-added distribution.

The simulations consider an environment in which each teacher teaches only one class per year. In this setting, σ_μ^2 and σ_θ^2 cannot be separately identified in the data, and only their sum, $V = \sigma_\mu^2 + \sigma_\theta^2$, is observed. We set $V = 0.05$, $\sigma_\varepsilon^2 = 0.25$, and $n = 20$, calibrated from the North Carolina elementary school data discussed in Appendix A.[18] We also fix the variance of the persistent component of value added at $\gamma_\mu = 0.03$. Holding V and γ_μ fixed implies that $V - \gamma_\mu = 0.02$ represents variation not captured by the persistent teacher effect, which may reflect either transitory teacher value added or contemporaneous classroom shocks.

Although Eq. (16) shows that the more inclusive forecast has a higher R^2 with teacher value added for any parameter values, the main focus of these exercises is how the prevalence of classroom shocks affects the relative performance of the two forecasts, and how this relationship varies across short and long panels. To capture this, we vary $p_\theta \in [0, 1]$, defined as the fraction of the non-persistent component $V - \gamma_\mu$ attributable to classroom shocks.[19]

[16] An important reason the leave-year-out forecast has a lower R^2 is the same source of variation it is designed to avoid. The leave-year-out information set is orthogonal to the contemporaneous classroom shock, but it still contains realized classroom shocks from the other years used to construct the forecast. Since these shocks are averaged over $T-1$ years rather than T years, they leave more noise in the leave-year-out information set than in the full-sample information set. This lowers the signal-to-noise ratio, leading to greater shrinkage and a forecast that is less correlated with value-added.

[17] Even if the goal were to rank teachers only on the persistent component of value added, the full-sample forecast would still have a higher R^2. For example, suppose value added is decomposed as $\mu_{jt} = \phi_j + \xi_{jt}$, where ϕ_j is a persistent teacher component with variance γ_μ and ξ_{jt} is a transitory component with variance $\sigma_\mu^2 - \gamma_\mu$. The R^2 with the persistent component can then be written as $R^2_{\phi|\hat{A}^{(k)}} = \gamma_\mu / \text{Var}(M^{(k)})$. Because the full-sample information set averages the transitory teacher component, classroom shocks, and student-level noise over more years, $\text{Var}(M^{\text{FS}}) < \text{Var}(M^{\text{LYO}})$, so $R^2_{\phi|\hat{A}^{\text{FS}}} > R^2_{\phi|\hat{A}^{\text{LYO}}}$.

[18] Gilraine et al. (2020) use similar parameter values in their simulations.

[19] $p_\theta = \sigma_\theta^2/(V - \gamma_\mu)$.

At $p_\theta = 0$, there are no classroom shocks, so $\sigma^2_\theta = 0$, $\sigma^2_\mu = 0.05$, and all non-persistent variation reflects transitory teacher value added. At $p_\theta = 1$, the opposite holds, so $\sigma^2_\theta = 0.02$, $\sigma^2_\mu = \gamma_\mu = 0.03$, and teacher value added is fully persistent across years.

Figure 2 reports the misclassification rate for teacher ranking using full-sample and leave-year-out forecasts. A teacher is misclassified in a given year if their forecast, $\hat{A}^{(k)}_{jt}$, places them in the bottom 5%, but their actual value added, μ_{jt}, does not. The top panel corresponds to $T = 3$, a short time frame often relevant for hiring, retention, or contract-renewal decisions in elementary school, and the bottom panel corresponds to $T = 10$. The horizontal axis varies p_θ from zero to one. Moving from left to right shifts variation away from transitory teacher value added and toward classroom shocks, while holding fixed both the total class-level variance and the persistent component of teacher value added.

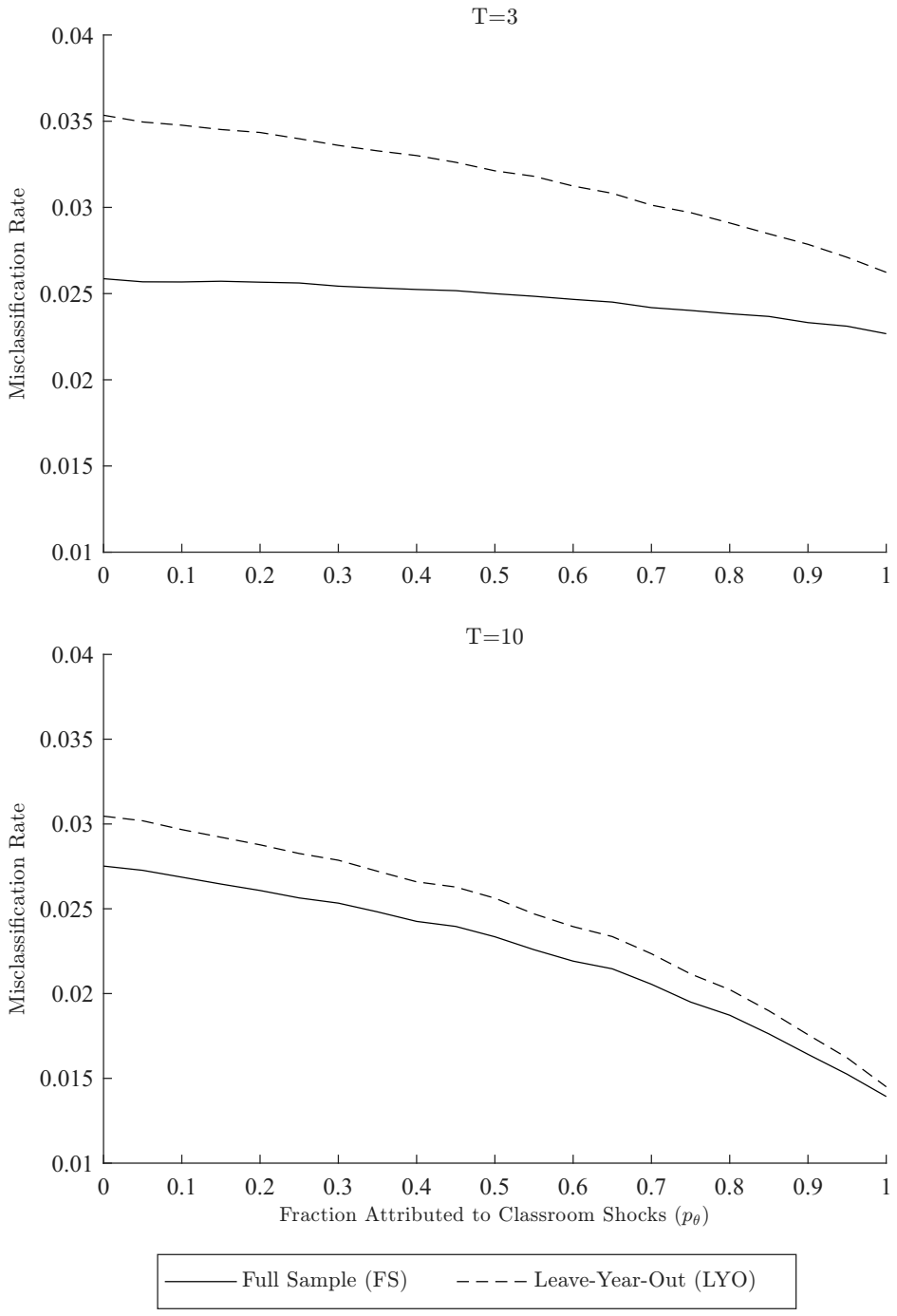
Consistent with Eq. (16), Figure 2 shows that the full-sample forecast yields uniformly lower misclassification rates than the leave-year-out forecast when identifying teachers in the bottom 5% of the value-added distribution. This holds even in the extreme case of $p_\theta = 1$, where classroom shocks account for all non-persistent class-level variation and teacher value added is fully persistent across years. Although across-year information is highly informative in this setting, removing an entire year of data discards valuable signal about the persistent teacher component. This loss outweighs the benefit of excluding contemporaneous classroom shocks. The cost of leaving out a year is especially large in short panels: when $T = 3$, the gap between the full-sample and leave-year-out forecasts remains substantial across all values of p_θ. When $T = 10$, the difference between the two forecasts becomes smaller because a single year represents a smaller share of the available information.

These results demonstrate that while there is some appeal to ranking teachers based on leave-year-out forecasts, which discard within-year information to reduce exposure to contemporaneous shocks, this approach leaves less signal about teacher value added in the information set. More inclusive forecasts retain more of this signal, producing stronger alignment with the underlying teacher effect. As a result, even when classroom shocks are substantial, using the full information set yields forecasts that are more closely aligned with μ_{jt} and produce more accurate rankings. For teacher ranking, the value of retaining an additional year of signal about teacher value added outweighs the benefit of orthogonality to contemporaneous classroom shocks.

5.2 Regression on Long-term Outcomes

We next examine how the trade-off between leave-year-out and more inclusive forecasts affects estimates of the relationship between teacher quality and long-term student outcomes.

Figure 2: Misclassification in Teacher Ranking



Note: The figure reports misclassification rates from Monte Carlo simulations in which teachers are classified as belonging to the bottom 5% of the value-added distribution using either full-sample (FS) or leave-year-out (LYO) forecasts. Misclassification error is defined as the probability that a teacher is classified in the bottom 5% based on the forecast but does not belong to the bottom 5% of the true value-added distribution.

These analyses use teacher value-added forecasts as measures of teacher quality and relate them to outcomes such as future test scores, course-taking, graduation, college attendance, or earnings. In this application, the trade-off becomes a bias-variance trade-off between two estimators, which can be assessed using a Hausman-type test.

Because different forecasts load on different test-score components, we begin by writing long-term outcomes as a function of the same components that enter test-score residuals. Let Y_i^* denote a long-term outcome for student i, who was taught in year t by teacher j. We write the long-term outcome as

$$Y_i^* = \beta^Y X_i + \kappa_\mu \mu_{jt} + \kappa_\theta \theta_{jt} + \kappa_\varepsilon \varepsilon_{it} + \nu_i.$$

The components μ_{jt}, θ_{jt}, and ε_{it} are the same components that enter the test-score residual in Eq. (1). The coefficients κ_μ, κ_θ, and κ_ε allow these components to have different effects on long-term outcomes.[20] The parameter of interest is κ_μ, the effect of teacher value added on the long-term outcome. The term ν_i captures remaining determinants of the long-term outcome that are not spanned by the components of test-score residuals and is orthogonal to these components by construction. Given an unbiased estimate of β^Y, define the residualized long-term outcome as

$$Y_i = Y_i^* - \hat{\beta}^Y X_i = \kappa_\mu \mu_{jt} + \kappa_\theta \theta_{jt} + \kappa_\varepsilon \varepsilon_{it} + \nu_i. \tag{17}$$

The parameters in Eq. (17) cannot be directly estimated because the components of test-score residuals are not separately observed. Empirical work instead uses a forecast $\hat{A}^{(k)}$ constructed from the information set $M^{(k)}$ as the measure of teacher quality, yielding the estimating equation

$$Y_i = \alpha + \kappa \hat{A}_{it}^{(k)} + \eta_i = \alpha + \kappa \left(w^{(k)} M_{it}^{(k)} \right) + \eta_i. \tag{18}$$

The estimand associated with Eq. (18), using information set $M^{(k)}$, is denoted by $\kappa^{(k)}$. It is a linear combination of the component-specific long-term effects, with weights determined

[20] For example, if ε_{it} contains unobserved student characteristics such as parental income, those characteristics may affect long-term outcomes differently than teacher value added.

by the relative loading of the forecast on each test-score component.[21]

$$\kappa^{(k)} = q_\mu^{(k)}\kappa_\mu + q_\theta^{(k)}\kappa_\theta + q_\varepsilon^{(k)}\kappa_\varepsilon,$$

where $q_\mu^{(k)} = w_\mu^{(k)}/w^{(k)}$, with $q_\theta^{(k)}$ and $q_\varepsilon^{(k)}$ defined analogously.

Since the leave-year-out forecast does not load on contemporaneous classroom shocks or student-level shocks, $q_\theta^{\text{LYO}} = 0$ and $q_\varepsilon^{\text{LYO}} = 0$, so $\kappa^{\text{LYO}} = \kappa_\mu$. Thus, using the leave-year-out forecast yields an unbiased estimand for the effect of teacher value added on long-term outcomes.

As a more inclusive alternative, we consider the leave-one-out forecast, which excludes student i's own test score from the information set but retains the other current-year observations. This exclusion has little effect on the forecast, since a single observation receives weight on the order of $1/(nT)$. In this case, $q_\varepsilon^{\text{LOO}} = 0$, but because the forecast treats part of the contemporaneous classroom shock as signal, any bias from using the leave-one-out forecast is determined by

$$\kappa^{\text{LOO}} - \kappa_\mu = q_\theta^{\text{LOO}}(\kappa_\theta - \kappa_\mu).$$

This expression shows that bias from the leave-one-out forecast is not automatic. Its size depends on both the weight placed on classroom shocks, q_θ^{LOO}, and the difference between the long-term effects of classroom shocks and teacher value added, $\kappa_\theta - \kappa_\mu$.

For the bias to be large, classroom shocks must account for a substantial share of the forecast signal and have long-term effects that differ meaningfully from teacher value added. These requirements are somewhat at odds. If classroom shocks account for a large share of the forecast signal, they are likely to reflect real learning inputs, in which case their long-term effects may be similar to those of teacher value added, so $\kappa_\theta \approx \kappa_\mu$. If instead classroom shocks reflect variation in measured achievement that does not correspond to learning, such as test-day disruptions that have little impact on human capital, then their long-term effects may differ from those of teacher value added, but their role in the forecast signal is likely to be limited, so q_θ^{LOO} would be small. In either case, the product $q_\theta^{\text{LOO}}(\kappa_\theta - \kappa_\mu)$, and therefore

[21] Since $\hat{\kappa}^{(k)}$ is the coefficient from a regression of Y_i on $w^{(k)}M_{it}^{(k)}$,

$$\begin{aligned}\operatorname{plim}\hat{\kappa}^{(k)} &= \frac{\operatorname{Cov}\left(Y_i, w^{(k)}M_{it}^{(k)}\right)}{\operatorname{Var}\left(w^{(k)}M_{it}^{(k)}\right)} \\ &= \frac{1}{w^{(k)}}\left[\kappa_\mu w_\mu^{(k)} + \kappa_\theta w_\theta^{(k)} + \kappa_\varepsilon w_\varepsilon^{(k)} + \frac{\operatorname{Cov}\left(\nu_i, M_{it}^{(k)}\right)}{\operatorname{Var}\left(M_{it}^{(k)}\right)}\right],\end{aligned}$$

with $\operatorname{Cov}\left(\nu_i, M_{it}^{(k)}\right) = 0$.

the bias from the leave-one-out forecast, would be small.

Although the leave-year-out forecast eliminates the potential bias from contemporaneous classroom shocks, it does so by discarding an entire year of information. This loss of information can make the estimator less precise, especially in short panels, and may make it harder to detect effects of teacher quality on long-term outcomes even when those effects are present. Holding fixed the residual variation in the long-term outcome regression, the variance of $\hat{\kappa}^{(k)}$ is inversely related to $\text{Var}(\hat{A}^{(k)})$. Since the leave-year-out forecast uses one less year of data than the leave-one-out forecast, a simple approximation is that the variance of the estimator is roughly $T/(T-1)$ times larger.[22]

In the long-term outcome setting, the leave-year-out forecast therefore yields a less precise estimator that eliminates potential bias from contemporaneous classroom shocks. The leave-one-out forecast yields a more precise estimator, but may be biased if the classroom shock component affects long-term outcomes differently than teacher value added. This creates a natural setting for a Hausman-type comparison. Under the null that the classroom shock component in the leave-one-out forecast does not bias the long-term outcome regression, both estimators recover κ_μ, and the difference between them should be small. Rejection of the null indicates that the classroom shock component introduces detectable bias, favoring the leave-year-out estimate despite its lower precision. Failing to reject the null should be interpreted more cautiously. It may indicate that the two estimates are close, or it may reflect limited precision, with their difference difficult to distinguish from sampling noise. In this case, the researcher must weigh the risk of retaining classroom-shock variation against the precision gains from including more data in the forecast.

We illustrate the bias-variance trade-off between leave-year-out and leave-one-out forecasts in long-term outcome regressions using Monte Carlo exercises calibrated to the North Carolina elementary school data discussed in Appendix A and used in the ranking analysis. In the ranking analysis, we varied the fraction of the non-persistent component, $V - \gamma_\mu$, attributable to classroom shocks. Here, we focus on the extreme case in which all non-persistent class-level variation is due to classroom shocks, so $\sigma^2_\theta = V - \gamma_\mu = 0.02$ and $\sigma^2_\mu = \gamma_\mu = 0.03$. This setting creates the greatest potential bias from using the leave-one-out forecast, since the classroom-shock component plays its largest possible role within the calibration. If classroom shocks are less prominent, the leave-one-out forecast becomes increasingly attractive because it retains more information while introducing less potential bias.

We focus on the short-panel case with $T = 2$.[23] In this setting, the leave-year-out

[22]The exact difference depends on the model parameters, though the leave-year-out forecast always yields a strictly larger variance than the leave-one-out forecast.

[23]As T increases, the two forecasts perform more similarly. The leave-year-out forecast becomes more precise because removing one year discards a smaller share of the data used to construct the forecast, while

forecast discards half of the data used to construct the forecast, while the leave-one-out forecast retains classroom shocks at their largest relative weight. Thus, combining a short panel with a calibration in which all non-persistent class-level variation is due to classroom shocks creates the setting in which the bias-variance trade-off between the two forecasts is most pronounced.

The long-term outcome equation is calibrated to the age-28 earnings specification in Chetty et al. (2014b). Specifically, we generate

$$Y_i = \bar{Y} + \kappa_\mu \mu_{jt} + \kappa_\theta \theta_{jt} + \eta_i,$$

where $\bar{Y} = 20{,}000$ and $\text{Var}(\eta_i) = (25{,}000)^2$. We set $\kappa_\mu = 2{,}000$ and vary the relative long-term effect of classroom shocks, $\kappa_\theta/\kappa_\mu \in [0, 1]$. At one extreme, $\kappa_\theta/\kappa_\mu = 1$ implies that classroom shocks have the same long-term effect as teacher value added, so including them in the forecast does not bias the estimate. At the other extreme, $\kappa_\theta/\kappa_\mu = 0$ implies that classroom shocks affect test scores but not long-term outcomes. In this case, classroom shocks are pure test-score noise with respect to Y_i, and including them in the forecast generates attenuation bias.

Figure 3 reports the Monte Carlo results for samples with $J = 20{,}000$, $J = 10{,}000$, and $J = 5{,}000$ teachers. In each panel, the horizontal axis varies κ_θ/κ_μ from zero to one, moving from the case in which classroom shocks affect test scores but not long-term outcomes to the case in which they have the same long-term effect as teacher value added.

The top panel reports the difference in root mean squared error between estimators based on the leave-one-out and leave-year-out forecasts, normalized by κ_μ:

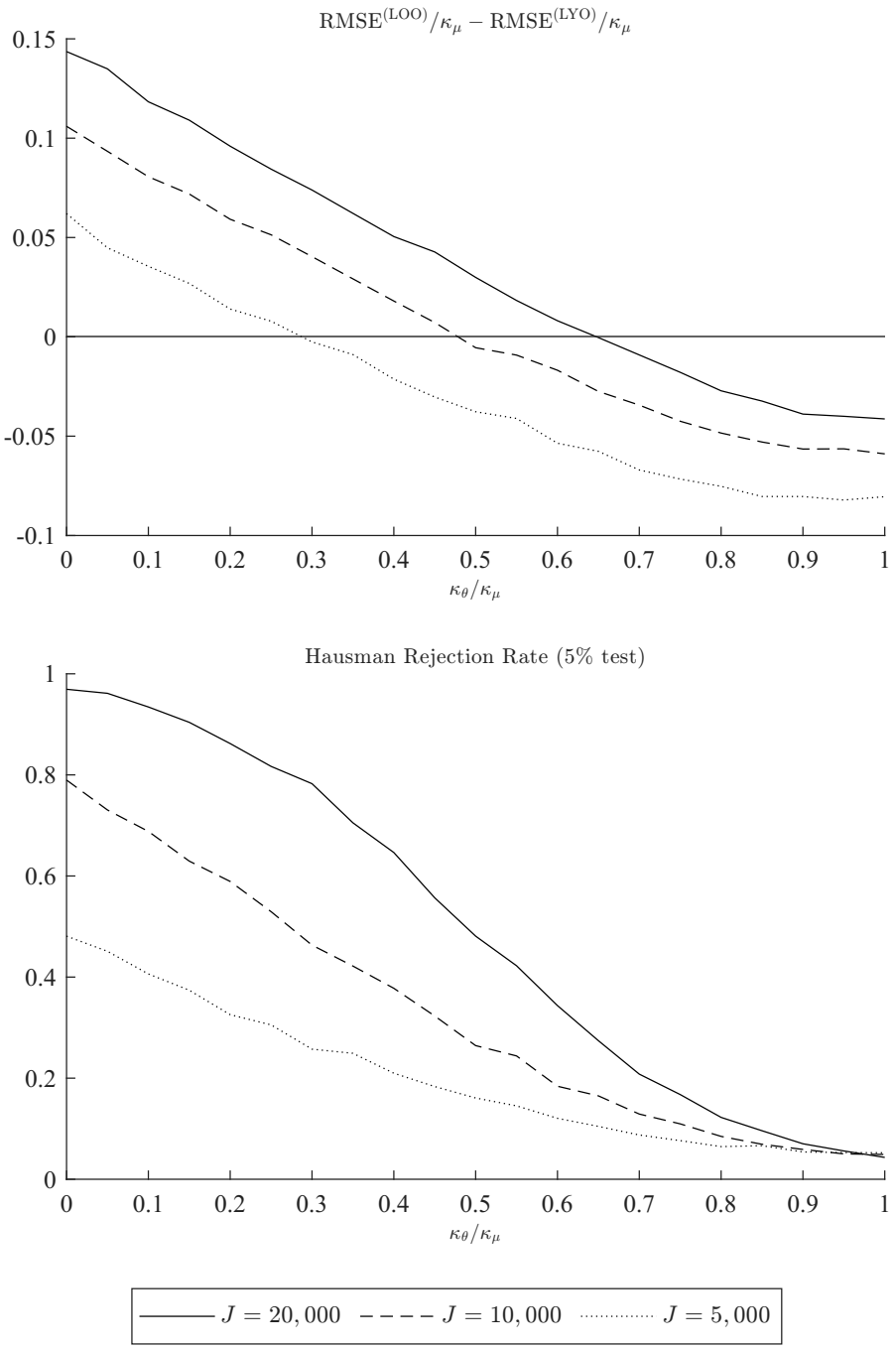
$$\text{RMSE}^{(\text{LOO})}/\kappa_\mu - \text{RMSE}^{(\text{LYO})}/\kappa_\mu.$$

Negative values indicate that the leave-one-out estimator has lower RMSE, while positive values indicate that the leave-year-out estimator has lower RMSE.

When $\kappa_\theta/\kappa_\mu = 0$, classroom shocks are pure test-score noise with respect to the long-term outcome. In this case, the leave-year-out forecast performs better because it removes this noise. For the largest sample, with $J = 20{,}000$, the leave-one-out estimator has an RMSE that exceeds the leave-year-out estimator's by roughly 15 percent of κ_μ. This advantage for the leave-year-out estimator is partly driven by the large sample size, which offsets the precision cost from discarding an entire year of test-score data. As the number of teachers falls, this precision cost becomes more important, and the RMSE difference between the two

the leave-one-out forecast becomes less exposed to classroom shocks, so q_θ^{LOO} declines.

Figure 3: Bias-Variance Trade-Off in Long-Term Outcome Regressions



Note: The figure reports Monte Carlo results for long-term outcome regressions using leave-one-out (LOO) and leave-year-out (LYO) forecasts. The top panel reports the difference in root mean squared error, normalized by κ_μ: $\text{RMSE}^{(\text{LOO})}/\kappa_\mu - \text{RMSE}^{(\text{LYO})}/\kappa_\mu$. Negative values indicate that the LOO estimator has lower RMSE, while positive values indicate that the LYO estimator has lower RMSE. The bottom panel reports rejection rates from a 5% Hausman test comparing the LOO and LYO estimates. The simulations vary κ_θ/κ_μ and report results for $J = 20{,}000$, $J = 10{,}000$, and $J = 5{,}000$ teachers.

estimators shrinks.

Moving to the right along the horizontal axis, the bias in the leave-one-out estimator declines as κ_θ/κ_μ approaches one. When classroom shocks have long-term effects similar to teacher value added, the classroom-shock component no longer creates meaningful bias, and the precision advantage of the leave-one-out forecast dominates. In this region, the leave-one-out estimator has lower RMSE than the leave-year-out estimator.

The bottom panel reports rejection rates from the Hausman test at the 5% level.[24] When $J = 20{,}000$, the test rejects in more than 95% of simulations when classroom shocks are pure test-score noise, indicating that the test is highly powered in this setting. With $J = 10{,}000$, the rejection rate falls to roughly 80% when $\kappa_\theta/\kappa_\mu = 0$. With $J = 5{,}000$, the test has more limited power, rejecting in roughly half of simulations.

The two panels show how the RMSE trade-off relates to the power of the Hausman test. Limited power arises in the same settings in which the leave-year-out estimator is least precise. These are also the settings in which the leave-one-out estimator offers the largest RMSE gains when classroom shocks have long-term effects similar to teacher value added. As J decreases, the RMSE profile shifts downward, indicating that the cost of incorrectly using the leave-one-out estimator falls while the potential gain from using it when it is unbiased rises. Thus, while detecting bias is more challenging when the Hausman test is underpowered, these are the same situations in which the precision gains from the leave-one-out estimator are most valuable.

6 Conclusion

This paper develops a framework for comparing value-added forecasts constructed from different information sets. By decomposing the forecast into the contributions of teacher value added, classroom shocks, and student-level shocks, the framework shows how the information set contributes signal about each component.

We use the framework to understand forecast calibration and clarify the interpretation of forecast-based regression tests. In general, these regressions compare an outcome to a model-implied forecast, so their coefficients show whether the forecast is correctly scaled for that outcome. For some test designs, when a suitable forecast is used, the coefficient can also illuminate particular features of the test-score process or the forecasts themselves. We illustrate this by showing how an aggregate forecast regression can be used to assess across-

[24] For the Hausman test, we compute $\text{Var}(\hat{\kappa}^{\text{LOO}} - \hat{\kappa}^{\text{LYO}})$ directly from the sampling distribution of the difference in the estimates across the Monte Carlo simulations. In empirical applications, this variance can instead be estimated from a joint estimation of the two specifications, which accounts for the covariance between $\hat{\kappa}^{\text{LOO}}$ and $\hat{\kappa}^{\text{LYO}}$.

teacher information in the test-score components. However, not all forecast-based regression tests can be interpreted in such a direct way. For the widely used teacher-switching test of Chetty et al. (2014a), we show that additional restrictions are required to interpret the coefficient as a test of teacher-level forecast bias.

We also apply the framework to study forecast choice and the advantages and disadvantages of leave-year-out and more data-inclusive forecasts in common applications. The central trade-off is that leave-year-out forecasts remove contemporaneous classroom shocks, but they also discard current-year teacher signal that may be useful in many applications. For teacher rankings, more inclusive forecasts perform better because the gain in teacher signal outweighs the cost of incorporating classroom shocks. For estimates of teachers' effects on long-term outcomes, leave-year-out forecasts may reduce bias, but more inclusive forecasts can provide substantial precision gains. These results show that the costs of bias and lost signal play different roles across applications. As a result, no single forecasting approach is likely to dominate in every application.

References

T. Ahn. The missing link: Estimating the impact of incentives on teacher effort and instructional effectiveness using teacher accountability legislation data. *Journal of Human Capital*, 7(3):230–273, 2013.

A. Bacher-Hicks, T. J. Kane, and D. O. Staiger. Validating teacher effect estimates using changes in teacher assignments in Los Angeles. Technical report, National Bureau of Economic Research, 2014.

B. Backes and M. Hansen. The impact of Teach For America on non-test academic outcomes. *Education Finance and Policy*, 13(2):168–193, 2018.

M. Bates, M. Dinerstein, A. C. Johnston, and I. Sorkin. Teacher labor market policy and the theory of the second best. *The Quarterly Journal of Economics*, 140(2):1417–1469, 2025.

R. Chetty, J. N. Friedman, and J. E. Rockoff. The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Working Paper 17699, National Bureau of Economic Research, December 2011.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632, 2014a.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9): 2633–79, 2014b.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers: Reply. *American Economic Review*, 107(6):1685–1717, 2017.

W. Delgado. Disparate teacher effects, comparative advantage, and match quality. *Economics of Education Review*, 106:102648, 2025.

M. Gilraine and O. McCarthy. The effect of the prior teacher on value-added. *Review of Economics and Statistics*, 108(3):645–662, 2024.

M. Gilraine and N. G. Pope. Making teaching last: Long-run value-added. Technical report, National Bureau of Economic Research, 2021.

M. Gilraine, J. Gu, and R. McMillan. A new method for estimating teacher value-added. Technical report, National Bureau of Economic Research, 2020.

E. A. Hanushek. Teacher deselection. In D. Goldhaber and J. Hannaway, editors, *Creating a New Teaching Profession*, pages 165–180. Urban Institute Press, 2009.

E. A. Hanushek and S. G. Rivkin. The distribution of teacher quality and implications for policy. *Annu. Rev. Econ.*, 4(1):131–157, 2012.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, 2015.

T. J. Kane and D. O. Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research, 2008.

C. Koedel. An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review*, 28(6):682–692, 2009.

C. Koedel, K. Mihaly, and J. E. Rockoff. Value-added modeling: A review. *Economics of Education Review*, 47:180–195, 2015.

J. A. Mincer and V. Zarnowitz. The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER, 1969.

North Carolina Department of Public Instruction. North carolina public school administrative data, 1997–2011, 2011. Distributed by the North Carolina Education Research Data Center, Duke University. https://childandfamilypolicy.duke.edu/north-carolina-education-research-data/ (accessed August 13, 2026).

J. Rothstein. Measuring the impacts of teachers: Comment. *American Economic Review*, 107(6):1656–84, 2017.

Appendix

A Data

We use an administrative dataset of North Carolina public-school students, teachers, and schools, provided by the North Carolina Education Research Data Center (NCERDC). Official records are collected by the North Carolina Department of Public Instruction (NCDPI). The panel contains student sociodemographic information as well as results from standardized End-of-Grade and End-of-Course exams for grades 3 to 12. Students are matched to test proctors, who may not be their regular classroom teachers.

We replicate the sample used in Rothstein (2017), restricting to students in grades 3 to 5 from 1997 to 2011, who are generally in self-contained classrooms. Contrary to many studies using the North Carolina data that exclude students attached to test proctors, we follow the sample construction procedure outlined in Rothstein (2017): each proctor-year who is not a classroom teacher is assigned a unique ID, so that student test scores are not used to infer the proctor's impact.

Table 5 compares our sample to that used in Rothstein (2017). The first three columns report the mean, standard deviation, and counts from Rothstein (2017); the next three report our reconstruction; the final column gives the Imbens-Rubin normalized difference (Imbens and Rubin, 2015). Seven of the ten statistics are trivially close ($|\text{norm. diff}| \leq 0.040$). Three differ. Class size differs (normalized difference 0.396): in the online appendix that contains the summary statistics table, Rothstein reports the figure as "not student weighted," counting students enrolled in each class, whereas our reconstruction counts students in the value-added estimation sample, which is smaller and yields a lower mean. The English language learner (ELL) and special-education shares are essentially the reverse of those in Rothstein (2017): Rothstein reports 0.085 and 0.023, while we obtain 0.021 and 0.086. Rothstein himself flags the high recorded ELL share as surprising. Other studies using the North Carolina data also report ELL shares are well below special-education shares (Ahn, 2013); the reversal appears to originate in how these two indicators are mapped from the raw NCERDC fields rather than in a difference in the underlying records. In any case, reversing the labels does not change any of the results of the analysis.

Table 6 compares the covariance structure of the test-score residuals across Rothstein (2017) (reported in the online appendix) with our sample. Rothstein estimates the covariance and correlation of classroom-average residualized achievement across different classes taught by the same teacher in different years, at lags 1 to 10. Because the paired classes are distinct, non-overlapping cohorts of students, persistence in class-average residuals across

Table 5: Sample comparison with Rothstein (Appendix Table A1)

	Rothstein (App. Table A1)			Our sample			
Variable	Mean	SD	N	Mean	SD	N	Norm. diff.
Class size	22.2[a]	5.0	357,036	20.2591	4.8037	356,840	0.396
No. subject-years per student	4.5	1.7	1,607,198	4.4953	1.6627	1,608,210	0.003
Test score (SD)	0.000	1.0	7,215,581	0.0393	0.9823	7,229,371	−0.040
Female	0.497		7,215,581	0.4963		7,229,371	0.001
Age (years)	10.5	0.9	7,213,590	10.5152	0.9482	7,229,371	−0.016
Free lunch eligible	0.449		3,926,246	0.4313		4,098,090	0.036
Minority (Black/Hispanic)	0.342		7,215,581	0.3425		7,229,371	−0.001
English language learner	0.085[b]		5,996,113	0.0214		6,009,903	0.286
Special education	0.023[b]		5,478,335	0.0857		6,009,903	−0.279
Repeating grade	0.014		7,215,581	0.0141		7,229,371	−0.001

Notes: Columns report means, standard deviations, and observation counts as printed in Rothstein's Appendix Table A1 (North Carolina column) and from the sample reconstructed here using Rothstein's replication code. Standard deviations are omitted for indicator variables. The normalized difference is $(\bar{x}_R - \bar{x}_S)/\sqrt{(s_R^2 + s_S^2)/2}$, where for indicators $s = \sqrt{p(1-p)}$; values below about 0.10 indicate negligible distributional differences, and values above about 0.25 indicate a meaningful difference (Imbens and Rubin, 2015). Significance tests are not reported because the two columns describe largely the same administrative records rather than independent samples.

[a] Rothstein labels class size "not student weighted." His figure counts students enrolled in each class; our reconstruction counts students retained in the value-added estimation sample, yielding a lower mean.

[b] Rothstein reports an English language learner share of 0.085 and a special-education share of 0.023; our reconstruction yields 0.021 and 0.086, essentially reversing the two. Rothstein notes the high ELL share as surprising. The reversal appears to arise in how these indicators are mapped from the raw data.

years cannot reflect repeated observations of the same students. However, it may still reflect persistent teacher effects, sorting, or other non-teacher inputs associated with the teacher's assignments.

Panel A reports, for each lag, the autocovariance of these residuals, its standard error, and the implied autocorrelation, following the layout of Rothstein's Appendix Table A2.

Panel B reports the within-year variance components: the total standard deviation of residuals, the individual (student-level) component, and the combined class-and-teacher component. Variances decompose additively; the table reports their square roots. Panel C then splits the combined class-and-teacher component into a teacher component and a residual classroom-shock component, reporting a lower-bound teacher standard deviation based on the lag-1 autocovariance and a quadratic estimate.

Across Rothstein's sample and ours, the estimated covariance structure is essentially identical: the largest absolute deviation across the table is 0.0009, within one unit of Rothstein's last reported digit, consistent with rounding and minor data-vintage differences.

Table 6: Covariance structure of test-score residuals: comparison with Rothstein (Appendix Table A2)

	English			Math		
	Rothstein	Our sample	Δ	Rothstein	Our sample	Δ
Panel A. Autocovariances and autocorrelations of residuals						
Lag 1	0.012	0.0121	+0.0001	0.032	0.0323	+0.0003
	(0.0002)	(0.00016)		(0.0002)	(0.00023)	
Autocorr.	0.359	0.3588	−0.0002	0.551	0.5514	+0.0004
Lag 2	0.011	0.0106	−0.0004	0.028	0.0280	0.0000
	(0.0002)	(0.00017)		(0.0003)	(0.00028)	
Autocorr.	0.317	0.3173	+0.0003	0.485	0.4849	−0.0001
Lag 3	0.009	0.0093	+0.0003	0.026	0.0255	−0.0005
	(0.0002)	(0.00019)		(0.0003)	(0.00033)	
Autocorr.	0.281	0.2812	+0.0002	0.442	0.4415	−0.0005
Lag 4	0.008	0.0083	+0.0003	0.023	0.0233	+0.0003
	(0.0002)	(0.00022)		(0.0004)	(0.00037)	
Autocorr.	0.250	0.2504	+0.0004	0.407	0.4073	+0.0003
Lag 5	0.008	0.0079	−0.0001	0.022	0.0219	−0.0001
	(0.0002)	(0.00024)		(0.0004)	(0.00042)	
Autocorr.	0.239	0.2397	+0.0007	0.384	0.3839	−0.0001
Lag 6	0.007	0.0071	+0.0001	0.021	0.0206	−0.0004
	(0.0003)	(0.00025)		(0.0005)	(0.00048)	
Autocorr.	0.218	0.2182	+0.0002	0.360	0.3595	−0.0005
Lag 7	0.007	0.0065	−0.0005	0.019	0.0192	+0.0002
	(0.0003)	(0.00028)		(0.0005)	(0.00053)	
Autocorr.	0.202	0.2029	+0.0009	0.333	0.3334	+0.0004
Lag 8	0.006	0.0064	+0.0004	0.018	0.0178	−0.0002
	(0.0003)	(0.00031)		(0.0006)	(0.00061)	
Autocorr.	0.201	0.2006	−0.0004	0.310	0.3100	0.0000
Lag 9	0.006	0.0058	−0.0002	0.017	0.0174	+0.0004
	(0.0003)	(0.00034)		(0.0007)	(0.00071)	
Autocorr.	0.184	0.1838	−0.0002	0.299	0.2996	+0.0006

Continued on next page

Table 6 — continued from previous page

	English			Math		
	Rothstein	Our sample	Δ	Rothstein	Our sample	Δ
Lag 10	0.006	0.0055	−0.0005	0.017	0.0167	−0.0003
	(0.0004)	(0.00037)		(0.0008)	(0.00079)	
Autocorr.	0.174	0.1746	+0.0006	0.285	0.2848	−0.0002
Panel B. Within-year variance components (standard deviations)						
Total SD	0.561	0.5606	−0.0004	0.544	0.5440	0.0000
Individual-level SD	0.542	0.5419	−0.0001	0.495	0.4954	+0.0004
Class + teacher-level SD	0.144	0.1436	−0.0004	0.225	0.2246	−0.0004
Panel C. Implied teacher standard deviation						
Lower bound (Lag 1)	0.110	0.1099	−0.0001	0.180	0.1796	−0.0004
Quadratic estimate	0.118	0.1182	+0.0002	0.192	0.1917	−0.0003

Notes: Each lag in Panel A reports the autocovariance of test-score residuals, its standard error (in parentheses), and the implied autocorrelation, reproducing the layout of Rothstein's Appendix Table A2 (North Carolina columns). "Rothstein" columns are as printed in that table; "Our sample" columns are computed here using Rothstein's replication code. Δ is the difference (our sample minus Rothstein) for the autocovariance and autocorrelation rows; it is omitted for the standard-error rows, which Rothstein reports to one significant figure. Panel B reports within-year variance components in standard-deviation units; Panel C reports the implied teacher standard deviation.

The largest absolute deviation across all entries is 0.0009, within one unit of Rothstein's last reported digit, so every cell matches to the precision he reports. Differences are shown in levels rather than normalized or scaled by the standard error: because the reported figures are rounded coarsely relative to their standard errors, scaling the level differences by those standard errors would convert rounding alone into apparent deviations of up to roughly two standard errors and misrepresent the agreement.

The reconstruction in Table 5 produces Rothstein's summary statistics closely, and the covariance-structure comparison in Table 6 matches to his reporting precision. Together, these indicate that our reconstruction recovers Rothstein's estimation sample, with the remaining differences attributable to variable definitions rather than to differences in the underlying data.