Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness*

Christine Mulhern Isaac M. Opper

April 7, 2022

Abstract

There is an emerging consensus that teachers impact multiple student outcomes, but it remains unclear how to measure and summarize the multiple dimensions of teacher effectiveness into simple metrics for research or personnel decisions. We present a multidimensional empirical Bayes framework and illustrate how to use noisy estimates of teacher effectiveness to assess the dimensionality and predictive power of teachers' true effects. We find that it is possible to efficiently summarize many dimensions of effectiveness and most summary measures lead to similar teacher rankings; however, focusing on any one specific measure alone misses important dimensions of teacher quality.

*We thank Michael Dinerstein, Andrew McEachin, Christopher Candelaria for many helpful discussions on how best to summarize teacher effectiveness. Eric Taylor and Ben Master provided helpful feedback on the paper. Seminar participants at the RAND Education and Labor Brownbag, Vanderbilt University, Princeton University, Amherst College, CESifo Economics of Education conference, IZA Workshop on the Economics of Education, AEA annual meeting, and AEFP also provided helpful comments. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A190148. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

I Introduction

Measuring teacher effects has been of longstanding importance in both research and policy. Accurately measuring teachers' impacts, often referred to as their value-added, is critical, as these measures are often tied to promotion and retention decisions, and value-added measures are used to answer a wide range of research questions.[1] A growing literature now documents that teacher effects extend beyond traditional measures of test score value-added, with teachers influencing outcomes such as student behavior, attendance, and grades (Gershenson (2016); Jackson (2018); Kraft (2019); Liu and Loeb (2019); Petek and Pope (2018)). Furthermore, traditional test score value-added does not necessarily identify the "best" teachers because teachers who effectively increase test scores are not necessarily effective at improving socio-emotional skills (Kraft (2019); Petek and Pope (2018)).

While there is an emerging consensus that teacher effects are multidimensional, it is not clear how to best measure teacher effectiveness or summarize the many dimensions of effectiveness into simple metrics that may be useful for personnel decisions. This paper discusses the challenges and implications of estimating teacher value-added in a multidimensional empirical Bayes framework, including important ways in which this is different from the single-dimensional model often used to estimate test score value-added. We also highlight complications that arise when using students' future outcomes, e.g., their test scores in the year after being taught, as one of the measures and provide a novel way to account for the fact that these outcomes are also affected by the students' future teachers. We then discuss approaches to construct summary measures of teacher effects, results on these summary measures and the implied dimensionality of teacher effects, and the practical importance of using these methods and results.

We consider two broad approaches for summarizing teacher effects which seek to balance the goal of identifying true effectiveness with practical evaluation limitations. First, we use

[1] For example, Dinerstein et al. (2021) use value-added estimates to measure human capital depreciation; Opper (2019) uses value-added measures to estimate endogenous peer effects; and Jackson and Bruegmann (2009) use value-added measures to estimate how teachers learn from each other.

principal component analysis to optimally reduce the dimensions of short-term effectiveness on which teachers are evaluated while minimizing information loss. Then we use the principal components to create summary measures of effectiveness and examine the dimensions of teacher effects. Second, we consider the case of a decisionmaker who wants to evaluate teachers based on their long-term effects, and construct summary measures which weight short-term measures of effectiveness so they optimally predict teachers' long-term effects.

While the two approaches are conceptually straightforward, implementing either one is complicated by the fact that each dimension of teacher effectiveness is estimated with noise. Furthermore, different measures may have different amounts of noise, and both the error with which each dimension of effectiveness is estimated and the true effects are correlated across the dimensions. Thus, we start by formally defining a multidimensional empirical Bayes framework to estimate teacher effects. While we are not the first to apply this framework for estimation of value-added models, the model are several important implications for estimation and interpretation that deserve discussion.

We discuss, for example, how the standard intuition that value-added measures are simply "shrunken" versions of the raw estimates breaks down in a multidimensional setting. Multidimensional empirical Bayes estimates for any given outcome will incorporate information about a teacher's estimated effect on all dimensions, since the estimated effectiveness on the other dimensions inform our belief about a teacher's true effect on the dimension in question.[2] For example, the best estimate of a teacher's impact on test scores will include information about the teacher's estimated impact on attendance. The magnitude and direction of the weights placed on the other dimensions, such as attendance, depend on the relative covariances of the true measures versus the error terms. Thus, even if teachers who increase attendance also tend to increase test scores, the multidimensional empirical Bayes estimates for test scores may put negative weight on the teachers' estimated effect

[2]In the single dimension setting, our prior is the mean, so teacher effects are shrunk to the mean. However, in the multidimensional setting, our prior is informed by our estimates of teacher effects on other dimensions and the covariances of these effects. So in this case, the extent to which we "shrink" or adjust our estimates depends on our estimates for other dimensions.

on attendance if the error terms are also positively correlated.

Next, we discuss how to use estimates from the multidimensional empirical Bayes framework to derive estimates of 1) the principal components, and 2) the relationship between the short-term effects and long-term effects. In this discussion, we explain why doing principal component analysis on the empirical Bayes estimates is different (and less preferred) than calculating the principal components of the true measures via an eigendecomposition of the estimated covariance matrix. We also show that multidimensional empirical Bayes estimates can be used as regressors to uncover the relationships between the true measures and outcomes of interest, a fact that is well-known in the single dimension case but is more nuanced in the multiple dimension case (Jacob and Lefgren (2008)).

We apply these techniques to estimate and summarize the effects of thousands of New York City teachers. Our estimates indicate that more than half of the variation in teacher effects on the outcomes we observe can be captured with one dimension. Furthermore, especially in elementary school, these short-term measures explain a large fraction of the teacher effects on students' long-term outcomes, around 80% for elementary school teachers. Our summary measures based on PCA are very similar to those which weight short-term effects based on their prediction of long-term effectiveness. In contrast, there is noticeable information loss when relying on a single outcome measure, rather than leveraging information from all of the measures. Likewise, there is little overlap in the teachers who are at the bottom five percent in terms of the summary measures and measures that rely on a single dimension. Thus, which measures are used for evaluation can have important implications for individual teachers.

This paper combines two important strands of the literature on teacher value-added. The first focuses on how to use imprecise measures of teachers' impacts on student test scores to evaluate teachers. This perspective led to the development of one-dimensional empirical Bayes estimation of teacher value-added (e.g., Kane and Staiger (2008); Chetty et al. (2014a)) and the design of teacher evaluation systems that aim to optimally combine teacher value-added measures with other measures of teacher practice, such as principal

ratings (e.g., Mihaly et al. (2013); Bacher-Hicks et al. (2020)). This strand, however, focuses exclusively on teachers' ability to improve students' test scores. More recent papers suggest that the focus on test scores may be insufficient, showing that teachers impact non-test score outcomes, that some teachers are better at improving non-test score outcomes than test score outcomes (and vice versa), and that teacher effects on non-test score outcomes are more predictive of teacher effects on students' long-term outcomes than teacher effects on test scores (e.g., Gilraine and Pope (2021); Gershenson (2016); Jackson (2018); Kraft (2019); Liu and Loeb (2019); Petek and Pope (2018)). Since they build on prior research, however, these papers generally separate the outcomes into traditional test score value-added and other measures, rather than focusing on how best to combine the various measures for evaluation or research.

While we are by no means the first to implement a multidimensional empirical Bayes framework, we hope to provide readers with a better understanding of the practical implications of using such a model, regardless of whether it is used to estimate teacher quality (e.g., Jackson (2018); Kraft (2019)), school quality (e.g., Beuermann and Jackson (2020); Abdulkadiroglu et al. (2020); Angrist et al. (2020)), hospital quality (e.g., Hull (2020)), or county effects (e.g., Chetty and Hendren (2018)). Much of the discussion about empirical Bayes estimates centers on them being "shrunken" versions of the raw estimates. While this is true in the single-dimension setting, in the multidimensional setting this intuition is no longer sufficient. We also discuss how to use this framework to conduct principal component analysis and assess the relationship of teacher effects on multiple outcomes, while accounting for the fact that the true measures of teacher effectiveness are unobserved. Specific to the teacher value-added context, we also develop a new approach to account for the fact that students' outcomes in the years after being taught by a teacher are also affected by the students' future teachers.

The paper proceeds as follows: Section 2 describes the multidimensional empirical Bayes framework and estimation; Section 3 describes the approaches for summarizing teacher effectiveness; Section 4 describes the data; Section 5 presents the results; Section 6 concludes.

II Multidimensional Empirical Bayes Estimation

This section starts by describing a multidimensional empirical Bayes estimation framework that can be used to estimate teacher effects when effects are multidimensional. In the single dimension, it is identical to prior models based on test score effectiveness (e.g., Kane and Staiger (2008)). The key difference is that we allow for a more complex variance structure. In particular, we allow for the error terms to be correlated across measures within a year and across years when the measures involve future outcomes. Allowing for the error terms to be correlated across years is important because different cohorts of students may have the same teachers in the future. We then briefly discuss the intuition behind the resulting empirical Bayes estimates, since they are no longer simply a shrunken version of the raw measures, and conclude by describing the steps for the full estimation procedure.

II.A Multidimensional Empirical Bayes Framework

Similar to other value-added papers, we start with a simple model for the production of student outcomes and role of teacher effects. We start by assuming there are K observed student outcomes of interest. These outcomes can include student test scores as well as other important outcomes such as students' attendance, behavior, self-efficacy, graduation rates, postsecondary attainment, and earnings, and can include outcomes measured in the year the teacher taught the student or in future years (Bacher-Hicks et al. (2019, 2020); Chamberlain (2013); Chetty et al. (2014a,b); Gershenson (2016); Gershenson et al. (2018); Gilraine and Pope (2021); Jackson (2018); Kraft (2019); Ladd and Sorensen (2017)). We denote student i's k^{th} outcome in year t as $y_{i,t,k}$ and let $X_{i,t}$ be a vector of student covariates that are not impacted by their teacher in year t. These characteristics generally include outcome measures from year $t-1$. The effect that teacher j would have on student i's k^{th} outcome in year t if she taught him is $\Theta_{j,t,k}$, which we refer to as her "true value-added" on the k^{th} measure. We assume that this effect is the same across students and hence do

not index $\Theta_{j,t,k}$ by i.[3]

Like most of the value-added literature, we assume that students' contemporaneous outcomes can be expressed as a linear function of: their teacher's effect on their outcome (which is assumed to be the same for all students); a vector of their covariates; a classroom-level shock shared by all students denoted as $\tilde{\nu}_{j,t,k}$; and an individual level shock denoted as $\epsilon_{i,t,k}$. Thus, our statistical model of student i's contemporaneous outcomes is:

$$y_{i,t,k} = \beta_k X_{i,t,k} + \Theta_{j,t,k} + \tilde{\nu}_{j,t,k} + \epsilon_{i,t,k} \tag{1}$$

When using future outcomes, such as the students' test scores or attendance in the year after being taught by teacher j, we need to tweak the above model slightly to account for an additional error term that is shared by some, but not all, of the students. Specifically, this error term accounts for the fact that many, but not all, of the students taught by teacher j also share the same classroom in future years. Their future outcomes are thus subject to a common shock, stemming from both the true value-added of their (shared) future teacher as well as the future classroom shock. We denote this term as $\phi_{j'(i),t,k}$, reflecting the fact that it depends on i's subsequent teacher j'.[4] The statistical model of student i's future outcomes is then:

$$y_{i,t,k} = \beta_k X_{i,t} + \Theta_{j,t,k} + \tilde{\nu}_{j,t,k} + \phi_{j'(i),t,k} + \epsilon_{i,t,k} \tag{2}$$

We next simplify our model by assuming that teacher effects on student outcomes are a combination of the teacher's persistent effectiveness and a year-specific shock to their effectiveness, i.e., $\Theta_{j,t,k} = \Theta_{j,k} + \eta_{j,t,k}$, for some persistent effect $\Theta_{j,k}$ and a year specific shock $\eta_{j,t,k}$. We do not incorporate drift in teacher effectiveness here as it complicates the model presentation, but the model can easily be extended to allow for drift. Appendix B

[3]See Delgado (2020) and Ahn et al. (2021) for work that relaxes this assumption. Note also that this assumption means we side-step the issues raised in Goldsmith-Pinkham et al. (2021).

[4]In this paper, we focus on outcomes in the year after being taught by teacher j. If instead we used outcome two years after a student is taught by teacher j, we should in theory include two ϕ error terms. However, as the outcomes become measured further out, the fraction of students who share future teachers likely decreases, so the future classroom shocks may be captured reasonably well by the individual error term.

shows that extending the model to account for drift in effectiveness does not change the interpretation of our results and Appendix H discusses how one can incorporate multiple years of data into the predictions both with and without drift in teacher effectiveness.

With this simplification, we define a new error term $\nu_{j,t,k} = \eta_{j,t,k} + \tilde{\nu}_{j,t,k}$, which we can think of as the error term that combines the classroom shock that is not caused by the teacher (embedded in the $\tilde{\nu}_{j,t,k}$ term) with the classroom-level shock that is caused by the teacher, but not related to a teacher's persistent effectiveness (embedded in the $\eta_{j,t,k}$ term).[5] We will not attempt to separate those two components of the error term in this paper, as it is not important for our research questions or most policy decisions.[6]

Dropping the k index to denote vectors of all the outcomes, the statistical model of student outcomes thus becomes:

$$y_{i,t} = \beta X_{i,t} + \Theta_j + \nu_{j,t} + \phi_{j'(i),t} + \epsilon_{i,t} \tag{3}$$

A key assumption is that the both the classroom and individual error terms are independently distributed across teachers and years, normally distributed, and have mean zero: $\nu_{j,t} \sim N(0, \Sigma_\nu)$ and $\epsilon_{i,t} \sim N(0, \tilde{\Sigma}_\epsilon)$. The assumption that the error terms are independently distributed across years means we assume that teachers are not consistently assigned to students who do worse (or better) than their covariates would suggest. This assumption is supported by several papers, which have tested this assumption for each of the measures we use (Chetty et al. (2014a); Bacher-Hicks et al. (2019); Petek and Pope (2018); Gilraine and Pope (2021)).

Although we assume the classroom and individual error terms are independent across years, we do not assume they are independent across measures within a year. This distin-

[5]For example, a dog barking outside the classroom during a test is a classroom-level shock not caused by a teacher, and a teacher getting sick on the day of an important lesson is a shock caused by the teacher but unrelated to persistent effectiveness.

[6]In contrast, you could imagine a principal who wants to reward some subset of teachers for their performance in the previous year, rather than to predict teacher performance in the subsequent year. In this case, separating the error terms would be important. As we discuss in Appendix B, separating them also is important when one allows for teacher drift.

guishes our model from others in the multidimensional teacher effectiveness literature which independently calculate value-added measures on each dimension (e.g., Jackson (2018); Petek and Pope (2018)). Further distinguishing our model is the error term $\phi_{j'(i),t}$ which is a function of student i' future teacher j', i.e., on the students' future teacher assignments. We assume this is also distributed normally, i.e., $\phi_{j'(i),t} \sim N(0, \Sigma_\phi)$, and independent across individuals who are taught by different teachers in future years.[7] We denote $\Sigma_\epsilon = \tilde{\Sigma}_\epsilon + \Sigma_\phi$, as the individual-level variance matrix that we identify using the approach described in the next section.[8]

In this formulation we account for errors in the raw value-added measures based on differential effectiveness of students' future teachers, and the fact that this is correlated across cohorts, but we do not directly control for the quality of the future teachers. From the Bayesian perspective, one could think of this as conditioning on the fact that there are future teachers of mixed effectiveness, but not on the specific identities of the students' future teacher. This is in contrast to other work that attempts to use measures of the subsequent teachers' effectiveness to adjust students' future outcomes (Candelaria et al., 2020; Gilraine and Pope, 2021). We opt against doing so here because a student's current teacher may impact their long-run outcomes by directing them to more effective teachers in subsequent years. In this case controlling for future teachers would inappropriately condition on an endogenous outcome. This is not without cost. Our method inadvertently advantages the teachers who happen to pass their students' off to more effective teachers in the same way it advantages teachers who happen to have students who do better than expected. If the measures are used for high-stakes decisions, whether this cost is worth the benefit is an important policy question.[9] Since the main focus in this paper is the variance of

[7]We also assume that $\phi_{j'(i),t} = 0$ for all contemporaneous outcomes. Thus, if the k^{th} measure is a contemporaneous outcome then both the k^{th} row and k^{th} column of Σ_ϕ consist of zeroes.

[8]As we discuss more later, Σ_ϕ is intimately connected with, and can be derived from, Σ_ν and the distribution of teachers' true value-added.

[9]That said, appropriately controlling for both teachers' quality without adding too much additional error is computationally challenging and requires teachers' students to move to many different teachers in the subsequent years (Jochmans and Weidner, 2019). The specifics of how best to do so is therefore an interesting research question.

and relationship between the true effects – and not on the prediction of individual teachers' effects for use in high-stakes settings – we opt for the approach that leads to consistent estimates of the covariance matrices.

We define a teacher's average residuals in year t as:

$$\theta_{j,t} = \frac{1}{N_j} \sum_{\forall i \in C(j,t)} y_{i,t} - \beta X_{i,t} \tag{4}$$

where $C(j,t)$ is the set of students teacher j teaches in year $t-1$ and $||C(j,t)|| = N_j$. From the statistical model and this definition, we get that:

$$\theta_{j,t-1} | \Theta_j \sim N\Big(\Theta_j, \Sigma_\nu + \zeta_j \Sigma_\phi + \frac{1}{N_j}\Sigma_\epsilon\Big) \tag{5}$$

where $\zeta_j \in [0,1)$ is determined by how many students of teacher j share the same teacher in the subsequent year. Specifically, if $N_{j,j'}$ is the number of students who have both teacher j and teacher j', then $\zeta_j = \sum_{\forall j'} \left(\frac{N_{j,j'}-1}{N_j}\right)^2$. The numerator is $N_{j,j'} - 1$ rather than $N_{j,j}$ because the individual component error term Σ_ϵ already contains the variance of future classroom shocks; $\zeta_j \Sigma_\phi$ thus reflects the additional within-classroom covariance of student residuals due to some of them sharing the same teacher in the subsequent year.

We further assume that teachers' true value-added is normally distributed with $\Theta_j \sim N(0, \Omega)$. Bayes Law then implies that:

$$\Theta_j | \theta_{j,t-1} \sim N\Big(\Omega_j^* \theta_{j,t-1}, \Sigma_j^*\Big) \tag{6}$$

where

$$\begin{aligned}
\Sigma_j &= \Sigma_\nu + \zeta_j \Sigma_\phi + \frac{1}{N_j}\Sigma_\epsilon \\
\Omega_j^* &= (\Sigma_j^{-1} + \Omega^{-1})^{-1} \Sigma_j^{-1} \\
\Sigma_j^* &= (\Sigma_j^{-1} + \Omega^{-1})^{-1}
\end{aligned}$$

While this provides the full posterior distribution under our normality assumptions, we generally focus on the mean of the posterior $\mathbb{E}[\Theta_j|\theta_{j,t-1}] = \Omega_j^*\theta_{j,t-1}$. We denote these empirical Bayes estimates as $\hat{\Theta}_j$.

The empirical Bayes framework relies on the strong assumption that both the true teacher effects and the error terms are normally distributed, which is a common assumption in the literature. More importantly, even if the normality assumption does not hold, the empirical Bayes estimates will be equivalent to the best linear predictors of the true teacher effects given the previous years' estimated teacher effects. See Appendix D for more discussion of this point and a proof that it is true in our setting.

II.B Intuition behind the Multidimensional Empirical Bayes Estimates

In the multidimensional setting, the matrix Ω_j^* contains the weights used to translate the various measures of teacher quality, $\theta_{j,t}$, into predictions of true teacher quality, Θ_j. In a single dimensional setting, this simply involves shrinking the measure of teacher quality toward the overall mean, where the shrinkage factor is based on the signal-to-noise ratio of the estimates. In the multidimensional setting, however, the translation from estimated measures to empirical Bayes estimates is more complicated. Most notably, unless both Σ_j and Ω are diagonal matrices, the empirical Bayes estimate of one dimension will incorporate information about the estimates of the other dimensions.[10] Intuitively, this makes sense as the estimated ability of the teacher to improve student test scores tells us something about the teachers true ability to increase student attendance.

To gain some insight into how information from other measures are incorporated into the empirical Bayes estimate of a particular measure we use a simple example with only two measures. Here, $\Omega = \begin{pmatrix} \sigma^2_{\Omega,1} & \rho_\Omega \\ \rho_\Omega & \sigma^2_{\Omega,2} \end{pmatrix}$ and $\Sigma_j = \begin{pmatrix} \sigma^2_{\Sigma,1} & \rho_\Sigma \\ \rho_\Sigma & \sigma^2_{\Sigma,2} \end{pmatrix}$. ρ_Ω and ρ_Σ correspond to the covariance between the two true measures of teacher effectiveness and the two error terms, respectively, rather than the correlation between the measures. If we denote $\Omega_j^* =$

[10] Technically, this statement is not quite true; if Σ_j is equal to Ω, the two forces pushing us to weight the other dimensions cancel each other out and the weights on the other dimensions is still zero. This is shown in the example below.

$\left(\begin{smallmatrix}\omega_{1,1} & \omega_{1,2}\\ \omega_{2,1} & \omega_{2,2}\end{smallmatrix}\right)$, we get:[11]

$$\omega_{1,1} = \frac{1}{det(\Omega + \Sigma_j)}\left[\sigma^2_{\Omega,1}\sigma^2_{\Omega,2} + \sigma^2_{\Omega,1}\sigma^2_{\Sigma,2} - \rho^2_{\Omega} - \rho_{\Omega}\rho_{\Sigma}\right] \tag{7}$$

$$\omega_{1,2} = \frac{1}{det(\Omega + \Sigma_j)}\left[\sigma^2_{\Sigma,1}\rho_{\Omega} - \sigma^2_{\Omega,1}\rho_{\Sigma}\right] \tag{8}$$

Thus, when calculating the empirical Bayes estimate of the first measure, the sign of the weight placed on the second measure depends on the relative covariances of the true measures versus the error term. This means the empirical Bayes estimate may put a negative weight on the second measure even when the two true measures are positively correlated if the error terms are even more positively correlated than the true measures.[12]

To understand why these negative weights may occur, it is important to recognize that the second measure of teacher effectiveness provides information on both the true teacher effects and the unobserved classroom quality, i.e., the error term. The estimate of the second measure of teacher effectiveness may be large either because: a) the teacher increased her students' second outcome, or b) the teacher got a good cohort of students who would have outperformed expectations regardless of their teacher. In case a) we should increase our estimate of the teacher's effect on the students' first measure, since the true measures being positively correlated imply that teacher who is good at increasing one outcome is also likely to be good at increasing the other outcome. On the other hand, in case b) we should decrease our estimate of her effect on the students' first measure, since the positive correlation of the error terms implies the class would likely outperform expectations on all outcomes even with an average teacher.[13] The relative variance and covariances of the true effects versus the error terms inform us whether a) or b) is the more likely explanation,

[11] See Appendix I for proof.

[12] While we use the term "correlated" here, equation (48) makes clear that the comparison of interest is actually a comparison of the weighted difference between the covariances rather than an unweighted comparison of the correlations.

[13] This analysis assumes that the true impact on the two measures is positively correlated, as is the error term for the two measures. In other words, it assumes both ρ_{Ω} and ρ_{Σ} are positive, which is what the data suggest.

and thus whether we should increase or decrease our estimate of the teacher's effect on her students' first measure after observing a high value of the second measure.

II.C Estimation Details

The multidimensional Bayesian approach outlined in Section II.A describes how to transform the estimated teacher residuals $\theta_{j,t}$ into the posterior estimates of the teacher's true value-added. In doing so, it considered the underlying hyperparameters, such as the covariance matrices of the true teacher value-added and the error terms, as fixed and known to the researcher. We now discuss how we estimate these hyperparameters from the data. While we outline the approach here, Appendix I contains the proofs that under the model and assumptions described above in Section II.A the approach outlined here provides consistent estimates of the relevant matrices.

We estimate the hyperparameters using the following six steps:

1. **Estimate $\hat{\beta}$ by fitting the OLS regressions at the student level with teacher fixed effects.** $y_{i,t} = \beta X_{i,t} + \nu_j$ where ν_j is a teacher fixed effect. Our vector of covariates, $x_{i,t}$ consists of indicators for gender, race, year, free and reduced-price lunch status, English language learner status, and cubic functions of previous outcomes.

2. **Estimate Σ_ϵ using the estimate of the error term from step 1.**

$$\hat{\Sigma}_\epsilon = \frac{1}{N} \sum_{\forall i} \hat{e}_{i,t} \hat{e}'_{i,t} \tag{9}$$

   where $\hat{e}_{i,t} = y_{i,t} - \hat{\beta} X_{i,t} - \hat{\nu}_j$ and $\hat{\nu}_j$ are the estimated teacher fixed effects.

3. **Calculate average teacher residuals using the estimates of $\hat{\beta}$ from step 1.** We estimate the teacher average residuals as:

$$\theta_{j,t} = \frac{1}{N_j} \sum_{\forall i \in C(j,t)} y_{i,t} - \hat{\beta} X_{i,t} \tag{10}$$

where $C(j,t)$ is the set of teacher j's students in year t and $N_j = ||C(j,t)||$.

In addition, for future outcomes we also calculate the average residuals for students who have teacher j in year t and teacher j' in year $t+1$ as:

$$\theta_{j,j',t} = \frac{1}{N_{j,j'}} \sum_{\forall i \in C(j,j',t)} y_{i,t} - \hat{\beta} X_{i,t} \tag{11}$$

where $C(j,j',t)$ is the set of teacher j's students in year t who have teacher j' in the next year and $N_{j,j'} = ||C(j,j',t)||$.

Finally, for future outcomes we also calculate average residuals for students who have teacher j in year t and do not have teacher j' in year $t+1$ as:

$$\theta_{j,-j',t} = \frac{1}{N_{j,-j'}} \sum_{\forall i \in C(j,-j',t)} y_{i,t} - \hat{\beta} X_{i,t} \tag{12}$$

where $C(j,-j',t)$ is the set of teacher j's students in year t who do not have teacher j' in the next year and $N_{j,-j'} = ||C(j,-j',t)||$.

4. **Estimate Ω using the cross-year covariance between the teacher average residuals.** We start by computing:

$$\hat{\Omega} = \frac{1}{J} \sum \theta_{j,t} \theta'_{j,t-1} \tag{13}$$

if there are J teachers. We use the covariance between teacher j's effects in time t on students who have teacher j' in the future and $t-1$ to avoid bias related to correlated errors across measures within a year.

For contemporaneous outcomes, this is enough to ensure valid estimates of Ω; for future outcomes, however, the teacher average residuals may be correlated across years because different cohorts of students shared the same teacher in subsequent

years. Specifically, if k and k' are both future outcomes, we get that:[14]

$$\frac{1}{J}\sum \theta_{j,t,k}\theta'_{j,t-1,k'} \rightarrow \Omega_{k,k'} + \overline{\zeta}\Omega_{\tilde{k},\tilde{k'}} \tag{14}$$

where $\tilde{k}$ and $\tilde{k'}$ are the indices that correspond to the same outcomes as k and k' measured in the concurrent year. For example, if $k = k'$ and corresponds to student test scores in the year after having teacher j, the cross-year covariance between a teacher's average residuals on that measure converges to the sum of the variance of the teacher effects on their students' future test scores and the variance of the teacher effects on their students' current test scores times $\overline{\zeta}$.

This suggests that it is possible to recursively estimate $\hat{\Omega}$ by first computing $\frac{1}{J}\sum \theta_{j,t}\theta'_{j,t-1}$ and then adjusting the terms corresponding to future outcome terms using the terms corresponding to the relevant current outcomes.[15] Instead, we estimate Ω for the future outcomes directly using:

$$\hat{\Omega}_{k.k} = \frac{1}{\tilde{J}}\sum_{\forall j,j'} \theta_{j,j',t,k}\theta_{j,-j',k',t-1} \tag{15}$$

where $\tilde{J}$ is the number of j, j' pairs in the data.[16]

5. **Back out $\Sigma_\nu + \overline{\zeta}\Sigma_\phi$ using excess variance in the within-year covariances.** We start by noting that $\mathbb{E}[\theta_{j,t}\theta'_{j,t}] = \Omega + \zeta_j\Sigma_\phi + \Sigma_\nu + \frac{1}{N_j}\Sigma_\epsilon$. We therefore estimate $\Sigma_\nu + \overline{\zeta}\Sigma_\phi$ as:

$$\hat{\Sigma}_\nu + \overline{\zeta}\hat{\Sigma}_\phi = \frac{1}{J}\sum \theta_{j,t}\theta'_{j,t} - \hat{\Omega} - \frac{1}{N_j}\hat{\Sigma}_\epsilon \tag{16}$$

[14]This assumes that we other measure future outcomes in the year after the students are taught by the teacher, rather than two or three years after. A similar relationship can be computed when considering outcomes more than one year in the future.

[15]This recursive approach is likely useful when considering outcomes more than one year after the students' are taught by the teacher. Another benefit of this approach is that it can be done without knowing who the students are linked to in the subsequent year, assuming that $\overline{\zeta}$ can be estimated.

[16]The fact that there are two approaches to estimate a single parameter hints at the fact that it is overidentified. One potential benefit of that is that the it would provide a way to empirically test the assumption that students' future teacher effectiveness is independent of their current teachers' effectiveness.

where J is the number of teacher-years and N_j is the number of students assigned to teacher j and $\overline{\zeta} = \frac{1}{J}\sum \zeta_j$.

6. **Optional: Separately estimate Σ_ν and Σ_ϕ using the within-year covariances of students who do not share the same teacher in the subsequent year.** Similar to above, we get that $\mathbb{E}[\theta_{j,j',t}\theta'_{j,-j',t}] = \Omega + \Sigma_\nu + \frac{1}{N_j}\Sigma_\epsilon$. We can estimate:

$$\hat{\Sigma}_\nu = \frac{1}{J}\sum \theta_{j,j',t}\theta'_{j,-j',t} - \hat{\Omega} - \frac{1}{N_j}\hat{\Sigma}_\epsilon \tag{17}$$

Given this estimate of $\hat{\Sigma}_\nu$ and the above estimate of $\Sigma_\nu + \overline{\zeta}\Sigma_\phi$, we can also back out an estimate of Σ_ϕ. This step is only required if one wants to incorporate variation in ζ_j across teachers in the construction of their empirical Bayes estimates in a similar ways as one incorporates variation in N_j into the empirical Bayes estimates.

After estimating each of the matrices above, we can compute the full error covariance matrix, i.e., $\hat{\Sigma}_j = \hat{\Sigma}_\nu + \zeta_j\hat{\Sigma}_\phi + \frac{1}{N_j}\hat{\Sigma}_\epsilon$, the optimal empirical Bayes weight matrix, i.e., $\hat{\Omega}^*_j = ((\hat{\Sigma}_j + \hat{\Omega})^{-1}\hat{\Omega}^{-1})'$, and the empirical Bayes estimates for each teacher, i.e., $\hat{\Theta}_j = \hat{\Omega}^*_j\theta_{j,t-1}$.

III Summarizing Teacher Effectiveness

Decision-makers often need to synthesize information on many dimensions of effectiveness into policy decisions and researchers frequently require low-dimensional measures of teacher effectiveness for their studies. Therefore, we discuss two natural approaches to summarize the multiple dimensions of teacher effectiveness into a more tractable measure of "effectiveness." First, we optimally summarize the short-term measures with a lower-dimensional vector, similar to principal component analysis (PCA). Second, we view the short-term measures as statistical surrogates for a long-run outcome of interest (e.g., teacher effects on high school graduation). While both approaches are conceptually straightforward, implementing each is complicated by the fact that we do not observe the true measures of teacher quality. When implemented correctly these approaches enable us to shed light on

interesting questions about the production of student outcomes.

III.A Summarizing the Short-Term Effects

The aim of our first approach is to reduce the vector Θ_j of teacher j's K measures of effectiveness into a smaller vector of H measures, while minimizing loss of information about teacher j's effectiveness.[17] This is essentially a principal component analysis (PCA), with one notable exception. Like in many applied settings, we do not observe the "true" measures we wish to summarize with PCA. Rather, our value-added estimates are noisy estimates of teacher's true effects and we need to determine how to account for this noise in our principal components analysis.

To account for this, we use the fact that the principal components correspond to the eigenvectors of the covariance matrix. Instead of using the covariance matrix of the empirical Bayes estimates, which an off-the-shelf PCA analysis would do, we account for the noise inherent in each estimate by instead using the covariance of the underlying true effects, i.e., Ω. Specifically, if Ω is full-rank it can be factorized into $W\Lambda W^{-1}$, where W is the matrix of right eigenvectors and Λ is a diagonal matrix of eigenvalues. The columns of W are then the principal components, ordered in importance by the value of the corresponding eigenvalue, with the amount of variation explained by a component being equal to the value of its corresponding eigenvalue divided by the sum of the eigenvalues.

In addition to optimally summarizing Θ_j, this factorization provides some interesting empirical results. In particular, it allows us to determine whether teacher effectiveness can be best summarized by teacher effects on "cognitive" vs. "non-cognitive measures" measures, as in Jackson (2018) and Petek and Pope (2018), or by teacher effects on "short-term" vs. "long-term" outcomes, as in Gilraine and Pope (2021). More generally, our approach allows us to better understand how well the various measures can be summarized and how the various measures align themselves when measures of teacher effects are restricted to two

[17] We provide a more formal definition of "minimizing loss of information" in Appendix E, where we provide more details about our approach.

or three dimensions.

Finally, while the eigenvalue decomposition gives rise to weights that optimally summarize teachers' true effects and allows us to construct these weights without observing each teachers' true effectiveness, we still need to construct an empirical Bayes estimate of the summary measure that results from these weights. If the optimal weights are denoted ω^*, the empirical Bayes estimate of $\Theta_j\omega^*$ is just $\hat{\Theta}_j\omega^*$, where $\hat{\Theta}_j$ is the empirical Bayes estimate of Θ. Thus, instead of re-estimating empirical Bayes estimates of the summary measures, we can just use the empirical Bayes estimates of the underlying measures and apply the optimal weights to this vector.[18] The resulting measures still differ from those one would obtain from conducting PCA on the empirical Bayes estimates directly (i.e., on $\hat{\Theta}$).[19] While subtle, we can think of this difference as the difference between "the best estimates of the best summary of Θ", rather than "the best summary of the best estimates of Θ."

III.B Relationship between Short-Term Measures and Long-Term Outcomes

Another natural approach for constructing summary measures is to view the short-term effectiveness measures as proxies, or statistical surrogates, for teacher effects on students' long-term outcomes, such as high school graduation or lifetime earnings (Petek and Pope (2018); Chetty et al. (2014b)). In this case, we care about the teacher's predicted effect on long-term outcomes given the teacher's short-term effects, which reduces the dimensions of effectiveness from the number of short-term outcomes to the number of long-term outcomes.

Formally, let $\tilde{\Theta}_j$ be the effect of teacher j on the long-run outcome of interest; for

[18]The fact that the empirical Bayes estimates of $\omega^*\Theta$ are $\omega^*\hat{\Theta}$ follows from the fact that if a $m \times 1$ vector x is distributed normally $N(\mu, \Sigma)$, then $w'x \sim N(w'\mu, w'\Sigma w)$ for any $m \times 1$ vector of weights w.

[19]That they give different results can be seen in the fact that the covariance matrix of the empirical Bayes estimates is $\Omega(\Omega + \Sigma_j)^{-1}\Omega$ rather than Ω. An empirical examination of the differences is in Table A.1.

simplicity, we assume there is a single long-term outcome of interest.[20] We then define:

$$\omega^* = \arg\min_{\omega} \frac{1}{J} \sum_{\forall j} (\tilde{\Theta}_j - \omega' \Theta_j)^2 \tag{18}$$

which means that $\omega^* = (\Theta'\Theta)^{-1}\Theta'\tilde{\Theta}$, where Θ is a $J \times K$ matrix where the j^{th} row is Θ_j' and $\tilde{\Theta}$ is a $J \times 1$ vector where the j^{th} row is $\tilde{\Theta}_j$.

These optimal weights depend on the true short-term and long-term effects, which we do not observe. One natural approach is to estimate ω^* by replacing the matrix of true effects with the matrix of empirical Bayes estimates, i.e., Θ with $\hat{\Theta}$, and replacing the true long-term effect with the estimated long-term effect, i.e., $\tilde{\Theta}_j$ with $\tilde{\theta}_j$. Doing so would mean the coefficients are defined by $\hat{\omega}^* = (\hat{\Theta}'\hat{\Theta})^{-1}\hat{\Theta}'\tilde{\theta}$

In the one-dimensional setting, it is well known that these estimated weights are asymptotically equivalent to the optimal weights, i.e., $\omega^* = \hat{\omega}^*$, (Jacob and Lefgren (2008)). However, it is not obvious from Jacob and Lefgren's (2008) proof in the single dimension that this result extends to the multidimensional framework, as their proof relies heavily on the fact that in the single dimension the empirical Bayes measures are simply a shrunken version of the raw measure, rather than a linear combination of multiple raw measures.

In Appendix F we prove that $\omega^* = \hat{\omega}^*$ in cases where Θ_j is multidimensional.[21] Thus, just as in the one-dimensional setting, researchers can use the multidimensional empirical Bayes estimates as covariates to uncover the relationship between the dependent variable and true teacher effects.

Importantly, this proof relies on the assumption that the same set of measures are used to estimate value-added as appear in the subsequent regression. For example, suppose that we observe math and ELA test scores and use these two measures to estimate value-added

[20] We could extend the results to when there are multiple long-run measures, but that would require us to determine how the various long-term measures should be weighted.

[21] Note that the proof also assumes that $\tilde{\theta}_j$ is equal to the true effect plus an error term and that the error term is uncorrelated with $\hat{\Theta}_j$. The assumption that the error term is uncorrelated with $\hat{\Theta}_j$ is likely wrong if the same cohort is used to estimate $\hat{\theta}_j$ as is used to estimate $\hat{\Theta}_j$. When using different cohorts, however, the assumption is similar to the one that underpins the value-added framework in Section II.A.

using the multidimensional empirical Bayes approach, before running three regressions: one that only includes the math value-added as a regressor; one that only includes ELA value-added as a regressor; and one that includes both. The proof in Appendix F shows that the estimated coefficients in the final regression, which includes both regressors, would converge to the same coefficients as would be obtained if the teachers' true value-added on math and ELA were observed and used as covariates. In contrast, it is not necessarily the case that the estimated coefficients from the first two regressions would converge to the same coefficients as those from regressions based on the true value-added measures. This is because both measures were used to compute the teachers' value-added measures, but only one measure appeared in the resulting regression. Similarly, if the value-added measures were independently estimated using a single dimension empirical Bayes framework, the coefficients on the first two regressions would be correct, in the sense that they would converge to the same coefficients as if the true measures were observed, but the coefficients from the last regression would not be correct. Thus, if researchers plan on using the value-added estimates as regressors in multiple regressions, they should estimate different value-added models depending on the set of measures they plan to use in the regressions. We provide some numerical examples in Appendix F, which illustrate that failing to use the appropriate value-added estimation approach can significantly bias the resulting coefficients.

This approach enables us to examine how teacher effects on various short-term outcomes are associated with long-term teacher effects. Furthermore, we are able to shed light on whether the observed short-term measures are sufficient to explain long-term teacher effects, or more precisely, the fraction of variation in long-term teacher effects explained by the observed short-term measures.[22] For this, we use the year-to-year covariance of the estimated long-term effect to estimate the true variance of the long-term effect, which we denote as $\hat{\sigma}^2_{LT}$, and the covariance of the true short-term effects (rather than the covariance of the empirical Bayes estimates).[23] Specifically, we compute the percent of the long-term

[22]This is intimately related to the R^2 measure from equation 18, but we cannot use the reported R^2 because we have noisy measures of both short-term and long-term teacher effects.

[23]This relies on the assumption that teachers are not consistently assigned students who do better or

variance explained by the true short-term measures as $\frac{\omega^{*'}\Omega\omega^{*}}{\hat{\sigma}^2_{LT}}$, where $\omega^{*'}$ are the coefficient estimates from equation 18.[24]

IV Data

We use anonymized administrative data from the New York City Department of Education (NYCDOE), which contain information on any student who attended grades 3-8 at a public school in New York City from the 2004-2005 school year until the 2013-2014 school year. We henceforth refer to school years using the spring year, e.g., the 2004-2005 school year as SY2005 or simply as 2005. The data contain yearly information about each student's grade-level, school attended, assigned math teacher, and assigned English teacher. They also contain some student demographic information, including the student's gender, whether the student is classified as an English Language Learner (ELL), and whether the student has been diagnosed with a learning disability.

We also observe students' year-end math and English (ELA) test scores, as well as the percent of days they attend school. Because tests change each year, we follow convention by normalizing these scores by subject, grade, and year to have a mean of zero and a standard deviation of one. To minimize the importance of outliers, we measure attendance by taking the log number of days absent, adding one to the number of absences to ensure we can take the log. We then multiply this by negative one so that positive values are preferred, as with the other outcomes. In addition, we observe the numeric grades that middle school students receive in all of their classes.

Since our focus is on teacher value-added, we drop students who are not matched to a

worse on their long-term outcomes than would be expected based on their observables. If there is sorting on the long-term measures, the true variance of the long-term effect will be biased upward. Assuming the short-term measures are unbiased, this would bias downward our estimates of how much variation in the long-term effects the short-term measures can explain.

[24]This formula stems from two results. First, for any weights ω, the variance of $\omega'\Theta_j$ is $\omega'\Omega\omega$, since the variance of Θ_j is Ω. Second, the weights on the short-term measures that best explain the long-term measure are ω^*, which stems both from the definition of a regression and the result that the coefficients from a regression using the value-added estimates as covariates can, loosely speaking, also be thought of as the coefficients from a regression using the true effects as covariates.

teacher in the data. In addition, we drop the students with all non-standard grade codes; most of these indicate separate special education classrooms, which are often exempt from the year-end tests, and it also removes students who are part of the Collaborative Teaching track. Together, these restrictions remove roughly 10% of the total observations. We also correct student-to-teacher matches that appear to be misclassifications. We re-code as missing any elementary school teacher who is assigned to more than 50 students or fewer than 5 students in a year. For middle school, we use an upper limit of 120 students a year. This only affects about 1.5% of the student-year observations.

Finally, since we require previous test scores to compute value-added measures, we cannot calculate value-added measures in the first year we observe data (SY2005), so this year is omitted from the analysis. Thus, our main analytic sample consists of students who attended and teachers who taught in public elementary and middle schools in New York City during 2006 to 2014. Table 1 provides summary statistics of our sample, which show that New York City is a very diverse district, with approximately 27% Black students, 38% Hispanic, and 18% Asian.

After restricting our sample, we observe approximately 20,000 teachers, about two-thirds of whom are in middle school. On average, teachers are in our data for about three years; the short time-span is largely due to limitations in the length of our panel, as the teachers on average have been teaching in New York City for over nine years.

We look at eight outcomes over which elementary school teachers' value-added can be constructed. These include math test scores, ELA test scores, future math and ELA test scores, attendance, future attendance, future math grades and future ELA grades. Most middle school teachers in our sample do not teach both math and English, so we focus on subject-specific outcomes.[25] This gives us six potential outcomes over which we construct teacher value-added: test scores in subject taught, future test scores, attendance, future attendance, grades in subject taught, and future grades in subject taught.

[25]We focus on middle school teachers in grades 6 and 7. We omit 8th grade teachers, since we do not have future test scores, attendance, or grades for their students.

V Results

V.A Empirical Bayes Estimates of Teacher Effectiveness

Figures 1 and 2 summarize the empirical Bayes estimates. We have standardized the outcomes at the student level, so a teacher who has a value-added estimate of 0.25 on some measure increases her students' outcomes by 0.25 student standard deviations on that measure. For most outcomes, there is meaningful variation in teacher effects. The main exception is attendance, for which there is very little variation in the empirical Bayes estimates.[26]

From the empirical Bayes estimates alone, it is impossible to know how much variation actually exists in teacher effectiveness. For example, it is unclear whether the limited variation in attendance value-added is because teachers do not affect student attendance or because measurement error means the empirical Bayes estimates are shrunk more for attendance than for other measures. To help shine light on this, Table 2 shows the estimated standard deviations of the true teacher effects rather than the empirical Bayes estimates. Note that although we do not observe the true teacher effects directly, their standard deviation is implied by Ω, which we can consistently estimate.

The results in Table 2 suggest there is little variation in how teachers impact student attendance (in both elementary and middle school). While there is a reasonable amount of variation across teachers in their average student attendance residuals (column (3) of Table 2), teachers with positive average student attendance residuals in one year are no more likely to have positive attendance residuals in the next year than a teacher with negative attendance residuals in the first year. This does not rule out the possibility that teachers have a big impact on their students' attendance; however, it suggests that their effectiveness on this metric varies more from one year to the next than other measures. This means that knowing a teacher's effect on student attendance in year $t-1$ is less helpful when predicting their overall effectiveness in year t than knowing their effect on other outcomes in year $t-1$.

Table 3 shows the correlation of teachers' true effects across our outcomes. Although we

[26]This could be due to loose tracking of attendance in the administrative data during our time span, however this would not explain why there is meaningful variation in the future attendance estimates.

cannot directly observe true effects, we can estimate the correlations as implied from our estimate of Ω. Effects on math and ELA tests are highly correlated, with a coefficient of 0.745. Teacher effects on current tests are also moderately correlated with effects on future tests. In elementary school, the correlation ranges from 0.11 to 0.39 while in middle school it is 0.88. Teacher effects on grades are also positively correlated with test score effects.

V.B The Dimensionality of Teacher Effectiveness

Next, we use principal component analysis to assess the dimensionality of teacher effects. Figure 3 shows the proportion of variance explained by each of the principal components (the values are reported in Table A.1). For elementary school, the first component explains 49% of the variation, and the first four components collectively explain over 93% of the variance. In middle school, the first component explains 62% of the variation and the first four components collectively explain 99% of the variance. These results indicate that our initial six or eight dimensions of effectiveness can be reduced to a smaller set of dimensions without losing much information. In both elementary and middle school, we focus on the first four components since they cumulatively explain over 90% of the variance and individually explain at least 5% of the variance.[27]

Table 4 and Figure 4 show the composition of the four main principal components in terms of the original outcomes. For elementary school, the first component is roughly a weighted average of all outcomes except for current attendance, which it does not weight at all. The second component primarily differentiates between teacher effects on current test scores, which are weighted positively, and their effects on future grades, which are weighted negatively. The third dimension roughly separates current effects on tests scores from future test score effects, and the fourth component appears to separate effects on math test scores

[27]The columns of Table A.1 highlight the importance of the methods for conducting principal component analysis. If we had instead conducted PCA on the empirical Bayes estimates we would overstate the importance of the first two components and understate the importance of the third through eighth ones. Conversely, if we had used the raw measures of teacher residuals we would have understated the variance explained by the first three components and overstated it for the 5th through 8th components. This stems in large part from the fact that the error terms for each outcome are less correlated within teachers than are teachers' true effects.

from ELA test score effects and future attendance.

For middle school teachers, the first component is primarily based on teacher effects on grades, though effects on test scores and attendance receive some positive weight. Component two separates teacher effects on test scores from effects on current grades. Component three appears to separate current effects on test scores and grades from future effects on attendance and grades, while component four appears to differentiate between future grades and effects on current grades and future attendance. Attendance receives very little weight in all of these components.

V.C The Relationship Between Short-term and Long-term Effects

Next, we look at how the empirical Bayes estimates of effectiveness relate to teachers' long-term effects on high school graduation.[28] Since our PCA results show that the variation in measures can be efficiently summarized by the first four components, we also run a regression using the first four components as covariates. Using the short-term measures as statistical surrogates for the long-term outcome of interest (Prentice (1989); Athey et al. (2019); Begg and Leung (2000)) is useful for thinking about how one might want to weight the dimensions of teacher effects when one primarily cares about long-term effects.

For this, we estimate each teacher's long-term impact on high school graduation for the students they taught in year t and regress that on the multidimensional empirical Bayes estimates of their short-term impact on students' outcomes (constructed using the students taught in year $t-1$).[29] Tables summarizing these regressions are in Table A.2. In general, they show that the first principal component of teacher effectiveness is the most predictive

[28]The Appendix contains results for long-term effects on earning a Regents diploma and advanced Regents diploma.

[29]Here, we use standardized measures of teacher effects, so the coefficients indicate the effect of a one standard deviation better teacher on dimension K, conditional on her effect on the other dimensions. While these results can in theory be used to assess the predictive power of individual measures of effectiveness, we encourage readers to instead think of them simply as indicative of a way to weight the short-term measures of effectiveness to create a summary measure. This is because the coefficients need to be interpreted as holding all other covariates fixed; for example, what is the impact of a teacher with a slightly higher impact on students' ELA scores while holding fixed their effect on students' math scores, future ELA and math scores, attendance, and future grades. This makes the interpretation complex and means the individual coefficients are estimated without much precision.

of high school graduation. For the individual empirical Bayes estimates, future attendance and future math grades are most predictive of high school graduation. The bottom rows of Tables A.2 and A.3 also report how much of the variation in long-term teacher effectiveness is explained by the short-term measures. Among elementary school teachers, roughly 70 to 80% of the variation in long-term effectiveness (in terms of high school graduation or regents diplomas) is explained by effectiveness on our short-term measures. This fraction is smaller for middle school teachers - for whom short-term effectiveness explains about twenty percent of the variation in long-term effectiveness - so the six middle school measures we use may be insufficient for assessing long term effects.

V.D Does the Method of Summarizing Effectiveness Impact the Ranking of Teachers?

We can use our results to construct summary measures of effectiveness. First, we construct a summary measure based on the weights implied by the first principal component. Second, we construct a summary measure based on the weights implied by the regression results described above. Given the high correlations between the various measures, we also conduct a third approach in which we first reduce the dimensionality using our principal component approach and then weight these components by their relationship with long-term measures.

Overall, our estimates of teacher effectiveness are similar across the different summary measures we consider. While the weights placed on individual measures vary across the approaches (see Figure 5 and Table A.4), the first three columns of Table 5 show that teachers ratings are highly correlated across the three weighting approaches, both in elementary and middle school. These summary measures are also correlated with value-added estimates based on teacher effects on the non-test outcomes and traditional test value-added, although less so than they are correlated with each other.[30] Table 5 and figures A.1 and

[30]The non-test value-added measures are averages of teacher effects on future grades, current attendance, future attendance, and current grades (among middle school teachers). For elementary school, test value-added takes the teacher's average effect on math and English tests, focusing only on the year the teacher has the student in her class. Middle school teachers' test value-added is based on test scores on the subject the teacher teaches.

A.2 show these correlations. Of note, the results suggest that the non-test measures alone are insufficient because incorporating test scores into the value-added estimation procedure improves the empirical Bayes estimates of non-test value-added. This is seen in the fact that the Multi-Dimension Non-Test VA measures are strongly correlated with the weighted summary measures, while the Single Dimension Non-Test VA measures are much less correlated.

Tables 6 and 7 also show the practical implications of evaluating teachers on various summary measures. Panel A shows the expected changes, in terms of high school graduation, test scores, and non-test outcomes, if the bottom five percent of teachers are replaced with an average teacher in terms of the relevant metric. Replacing the bottom five percent of elementary school teachers based on the summary measures is associated with approximately 14pp higher high school graduation rates, relative to 8pp if decisions are based on test score value-added and 11pp for non-test value-added. Naturally, if the goal is instead to improve test scores, test score value-added will have the largest impact and similarly with non-test outcomes.

Panel (B) in Tables 6 and 7 shows the overlap in teachers who are in the bottom five percent on each of the metrics. In general, which measure is used for evaluation will have different implications for individual teachers. In elementary school, there is relatively little overlap between who is in the bottom five percent on the summary measures and who is in the bottom for test score and non-test value-added. For example, among elementary school teachers in the bottom five percent on test score value-added, only 17% are in the bottom for non-test value-added and 37% are in the bottom for the eigenvalue summary measure. Among middle school teachers, there is higher overlap (75-86%) in terms of the bottom five percent for the summary measures and the non-test outcomes, but less overlap with test score value-added. Thus, while different measures of teacher effectiveness are highly correlated, which measure is used for evaluation purposes can have important consequences for long-term outcomes and for which teachers are affected by personnel decisions.

It is worth emphasizing, however, that the results here are all in a context in which no in-

centives are attached to attendance and grades. One natural concern is that that grades and attendance measures are more gameable than test scores, since they are generally recorded directly by the teacher. Although using future attendance and grades might alleviate that issue to some extent, doing so may complicate the intra-school dynamics since it would imply that a 4^{th} grade teacher's evaluation would depend on the 5^{th} grade teacher's subjective evaluation of their students. This may have unintended consequences and we believe that districts would, understandably, be hesitant to introduce a high-stakes evaluation that relies on students' grades. Even setting these issues aside, relying on future measures requires the principal to wait additional years before being able to measure teacher effectiveness, delaying feedback and reducing the amount of information available at the time decisions are made. In addition, school districts may not have access to all of measures we consider "non-test score measures" or, alternatively, may have access to additional measures. Despite these challenges, the results described above suggest that test scores alone do not sufficiently summarize teacher effectiveness. Thus, developing non-gameable measures that adequately summarize teacher effectiveness would be quite valuable and is an important next step in the research on teacher evaluation.

Finally, it is important to consider the implications of missing data. As the number of measures used in teacher evaluations grows, it is more likely that some set of teachers will be missing outcomes on at least one dimension. While it is still possible to compute summary measures in the presence missing data, the distribution of these measures will vary based on the number of measures observed for a given teacher.[31] Importantly, in the empirical Bayes framework, teachers who are missing more measures will be more likely to be rated as average because their estimates will be shrunk more towards the mean. Thus, a teacher's likelihood of being in the bottom or top of the distribution – and perhaps subject to rewards or punishments – will depend on the number of measures observed. Concretely, when we randomly remove half of the observed measures from a subset of the teachers and

[31]In the multidimensional empirical Bayes framework, one can use the information about a teacher's estimated effects on the measured dimensions and estimated covariance of teacher effects across all dimensions to compute measures of teacher effectiveness even if the teacher is missing data for some outcomes,

re-estimate their value-added, their chance of being in the bottom 5% based on the summary measure decreases from around 6% to less than 2%.[32] Thus, any policy applications of these measures should carefully consider how missing data are treated and fairness and efficiency implications from using these measures in personnel decisions.

VI Conclusion

Accurately measuring the multiple dimensions of teacher effectiveness is important given growing evidence that teacher effects are multidimensional and the use of value-added in personnel decisions. Furthermore, it is important to figure out how to efficiently combine multiple measures of effectiveness into summary measures that can be used for policy and personnel decisions. Creating summary measures of effectiveness is, however, complicated by the fact that teacher effects are measured with noise, some outcomes are unobserved, and the error with which teacher effects are measured is correlated across outcomes.

This paper walks through the process and implications of estimating teacher value-added in a multidimensional framework. We show that, in a multidimensional setting, empirical Bayes estimates are not simply shrunken estimates of the raw versions, since they incorporate information about effectiveness on other dimensions. In addition, the multidimensional setting has implications for conducting principal component analysis and using the empirical Bayes estimates as covariates. The methods used to compute empirical Bayes estimates also influence estimates of the dimensionality of teacher effects and rankings of teachers.

Using data on New York City elementary and middle school teachers, we show that much of the variation in teacher effects, and their impacts on long-term outcomes, can be explained in a single dimension of effectiveness. We explore three approaches for summarizing teacher effectiveness, and all three measures lead to similar rankings of teachers. However, these

[32] Also of note, the fact that they have a 6% chance of being in the bottom 5% is itself a function of the fact that we focused on teachers with the complete set of measures and they are more likely to be in the tail of the effectiveness distribution than teachers with fewer measures.

summary measures are only moderately correlated with traditional test score value-added, and there is little overlap between teachers who are at the bottom five percent in terms of the summary measures and test score value-added.

Although our focus has been on the teacher setting, there are numerous other examples where researchers or policymakers want to efficiently summarize noisily estimated multidimensional effects. These include measuring hospital or physician effectiveness, employee productivity, and location-specific effects. All of these contexts, including the teacher setting, have complications that can make policy implementation complex. It is beyond the scope of this paper to examine, for example, how changing value-added measures may impact teacher incentives and effectiveness. We also assume that effects were consistent across individuals, that none of the measures were biased, and that all measures were continuous and normally distributed. Many of these complications have been studied in single-dimensional settings, e.g., Dinerstein and Opper (2020); Hull (2020); Delgado (2020); Angrist et al. (2017); Gilraine et al. (2020). A natural extension is therefore to combine our results on how to fairly measure and summarize noisily estimated multidimensional effects with approaches for dealing with these complications. In doing so, we can implement more efficient and fair personnel policy across a range of contexts.

References

Abdulkadiroglu, Atila, Parag A. Pathak, Jonathan Schellenberg, and Christopher Walters, "Do Parents Value School Effectiveness," *American Economic Review*, 2020, *110* (5), 1502–1539.

Ahn, Tom, Esteban M. Aucejo, and Jonathan James, "The Importance of Matching Effects for Labor Productivity: Evidence from Teacher-Student Interactions," 2021.

Amemiya, Yasuo, "Instrumental variable estimator for the nonlinear errors-in-variables model," *Journal of Econometrics*, 1985, *28* (3), 273–289.

Angrist, Joshua D., Peter Hull, Parag A. Pathak, and Christopher Walters, "Simple and Credible Value-Added Estimation Using Centralized School Assignment," 2020.

_ , _ , Parag Pathak, and Christopher Walters, "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, 2017, pp. 871–919.

Athey, Susan, Raj Chetty, Guido W. Imbens, and Hyunseung Kang, "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," 2019.

Bacher-Hicks, Andrew, Mark J. Chin, Heather C. Hill, and Douglas O. Staiger, "Explaining Teacher Effects on Achievement Using Commonly Found Teacher-Level Predictors," 2020.

_ , _ , Thomas J. Kane, and Douglas O. Staiger, "An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys," *Economics of Education Review*, 2019, *73*, 101919.

Begg, Colin B and Denis HY Leung, "On the use of surrogate end points in randomized trials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2000, *163* (1), 15–28.

Beuermann, Diether W. and C. Kirabo Jackson, "The Short and Long-Run Effects of Attending The Schools that Parents Prefer," *Journal of Human Resources*, 2020.

Candelaria, Christopher A., Brendan Bartanen, and Sy Doan, "Rethinking Value-Added: Medium-Term Teacher Effects on Student Achievement," 2020.

Chamberlain, Gary E., "Predictive effects of teachers and schools on test scores, college attendance, and earnings," *Proceedings of the National Academy of Sciences*, 2013, *110* (43), 17176–17182.

Chetty, Raj and Nathaniel Hendren, "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates," *The Quarterly Journal of Economics*, 2018.

_ , John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 2014, *104* (9), 2593–2632.

_ , _ , and _ , "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 2014, *104* (9), 2633–2679.

Delgado, William, "Heterogeneous Teacher Effects, Comparative Advantage, and Match Quality: Evidence from Chicago Public Schools," *Working Paper*, 2020.

Dinerstein, Michael and Isaac M. Opper, "The Effect of Value-Added Incentives on Multidimensional Teacher Output: Evidence from Tenure Reform in New York City," 2020.

_ , Rigissa Megalokonomou, and Constantine Yannelis, "Human Capital Depreciation," 2021.

Gershenson, Seth, "Linking Teacher Quality, Student Attendance, and Student Achievement," *Education Finance and Policy*, 2016, *11* (2), 125–149.

_ , Cassandra M. D Hart, Joshua Hyman, Constance Lindsay, and Nicholas W Papageorge, "The Long-Run Impacts of Same-Race Teachers," Working Paper 25254, National Bureau of Economic Research November 2018.

Gilraine, Michael and Nolan G. Pope, "Making Teaching Last: Long-Run Value-Added," 2021.

_ , Jiaying Gu, and Robert McMillan, "A New Method for Estimating Teacher Value-Added," *NBER Working Paper*, 2020.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesar, "On Estimating Multiple Treatment Effects with Regression," 2021.

Hausman, Jerry A., Whitney K. Newey, Hidehiko Ichimura, and James L. Powell, "Identification and estimation of polynomial errors-in-variables models," *Journal of Econometrics*, 1991, *50* (3), 273–295.

Hong, Han and Elie Tamer, "A simple estimator for nonlinear error in variable models," *Journal of Econometrics*, 2003, *117* (1), 1–19.

Hu, Yingyao and Susanne M. Schennach, "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 2008, *76* (1), 195–216.

Hull, Peter, "Estimating Hospital Quality with Quasi-Experimental Data," 2020.

Jackson, C. Kirabo, "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes," *Journal of Political Economy*, 2018, *126* (5), 2072–2107.

_ and Elias Bruegmann, "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics*, 2009.

Jacob, Brian A. and Lars Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluations in Education.," *Journal of Labor Economics*, 2008, *26* (1), 101–136.

Jochmans, Koen and Martin Weidner, "Fixed-effect regressions on network data," *Econometrica*, 2019, *87* (5), 1543–1560.

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

Kraft, Matthew A., "Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies," *Journal of Human Resources*, 2019, *54* (1), 1–36.

Ladd, Helen F. and Lucy C. Sorensen, "Returns to Teacher Experience: Student Achievement and Motivation in Middle School," *Education Finance and Policy*, 2017, *12* (2), 241–279.

Lewbel, Arthur, "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 1998, *66* (1), 105–121.

Liu, Jing and Susanna Loeb, "Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School," *Journal of Human Resources*, 2019.

Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger, and J.R. Lockwood, "A Composite Estimator of Effective Teaching," 2013.

Opper, Isaac M., "Does Helping John Help Sue? Evidence of Spillovers in Education," *American Economic Review*, March 2019, *109* (3), 1080–1115.

Petek, Nathan and Nolan G. Pope, "The Multidimensional Impact of Teachers on Students," *Working Paper*, 2018.

Prentice, Ross L, "Surrogate endpoints in clinical trials: definition and operational criteria," *Statistics in medicine*, 1989, *8* (4), 431–440.

VII Tables and Figures

Table 1: Summary Statistics

	Elementary School		Middle School	
	Mean	SD	Mean	SD
(A) Student Demographics				
Asian	0.18	0.38	0.17	0.38
Black	0.26	0.44	0.28	0.45
Hispanic	0.39	0.49	0.38	0.49
White	0.16	0.37	0.15	0.36
Male	0.49	0.50	0.49	0.50
English Language Learner	0.12	0.32	0.10	0.30
Free or Reduced Price Lunch	0.79	0.40	0.80	0.40
(B) Student Achievement				
Math Test Score	0.01	1.00	0.00	1.00
English Test Score	-0.00	1.00	0.02	1.00
Ln(Days Absent + 1)	1.94	0.95	2.07	0.99
Math Grade	79.66	11.30	80.46	12.02
English Grade	78.63	10.59	79.14	11.32
(C) Teachers				
Years Teaching at Current School	7.14	5.87	5.57	5.33
Years Teaching in the District	9.35	6.78	7.85	6.60
Male	0.14	0.34	0.24	0.43
(D) Counts				
Teachers	7,077	0	13,896	0
Teacher-Years	20,707	0	48,593	0
Teacher-Subject-Years	20,868	0	52,312	0
Students	182,807	0	617,921	0
Student-Years	477,286	0	1,482,360	0
Student-Subject-Years	477,286	0	2,692,055	0

Notes: Column 1 shows the mean for elementary school teachers and students. Column 2 shows the standard deviation for elementary school teachers and students. Column 3 shows the mean for middle school teachers and students. Column 4 shows the standard deviation for middle school teachers and students. Elementary school is defined as 5th grade and middle school is defined as 6th - 7th grade. The means and standard deviations are weighted by the frequency with which students and teachers appear in the sample. Test scores are standardized at the student level prior to restricting the sample. The sample is restricted to teachers with at least ten (tested) students. Data includes students and teachers from the 2005-06 school year through the 2013-14 school year.

Table 2: Standard Deviations of Teacher Effects

	True Measures (1)	Empirical Bayes (2)	Raw Measures (3)
(A) Elementary School			
Math Test	0.204	0.153	0.314
ELA Test	0.162	0.112	0.286
Attendance	0.010	0.004	0.055
Future Math Test	0.102	0.059	0.291
Future ELA Test	0.170	0.108	0.264
Future Attendance	0.188	0.082	0.274
Future Grades Math	0.153	0.122	0.404
Future ELA Grades	0.175	0.077	0.371
(B) Middle School			
Test Scores	0.177	0.126	0.271
Attendance	0.014	0.009	0.050
Current Grade	0.314	0.218	0.462
Future Test Scores	0.164	0.123	0.329
Future Attendance	0.139	0.102	0.322
Future Grade	0.217	0.151	0.490

Notes: Column 1 reports the standard deviation of true teacher effects based on the covariance matrix Ω^*. Column 2 reports the standard deviation of the empirical Bayes estimates of teacher effects. These estimates understate the true standard deviation of teacher effects. Column 3 reports the standard deviation of the raw estimates of teacher effects (i.e. their average student residuals). These estimates overstate the true standard deviation of teacher effects. Panel (A) is for elementary school, defined as 5th grade. Panel (B) is based on middle school, defined as grades 6th-7th. For middle school, test score and grade value-added are only for the one subject a teacher teaches. Elementary school teachers teach both math and English. The units for all measures are the standard deviations of student outcomes.

Table 3: Correlation of Teacher Effects on Various Outcomes

	(A) Elementary School							
	Math Test Scores (1)	ELA Test Scores (2)	Attendance (3)	Future Math Test (4)	Future ELA Test (5)	Future Attendance (6)	Future Math Grades (7)	Future ELA Grades (8)
Math Test	1.000	0.745	0.097	0.107	0.389	0.059	0.331	0.145
ELA Test	0.745	1.000	0.078	0.164	0.331	0.055	0.500	0.173
Attendance	0.097	0.078	1.000	-0.282	-0.114	0.055	-0.090	-0.043
Future Math Test	0.107	0.164	-0.282	1.000	0.353	0.532	0.333	0.337
Future ELA Test	0.389	0.331	-0.114	0.353	1.000	0.440	0.799	0.466
Future Attendance	0.059	0.055	0.055	0.532	0.440	1.000	0.378	0.907
Future Grades Math	0.331	0.500	-0.090	0.333	0.799	0.378	1.000	0.451
Future ELA Grades	0.145	0.173	-0.043	0.337	0.466	0.907	0.451	1.000

	(B) Middle School					
	Test Score (1)	Attendance (2)	Grade in Subject (3)	Future Test Score (4)	Future Attendance (5)	Future Grade in Subject (6)
Test Scores	1.000	0.172	0.128	0.875	0.244	0.233
Attendance	0.172	1.000	0.080	0.187	0.328	0.063
Current Grade	0.128	0.080	1.000	0.125	0.094	0.789
Future Test Scores	0.875	0.187	0.125	1.000	0.600	0.376
Future Attendance	0.244	0.328	0.094	0.600	1.000	0.456
Future Grade	0.233	0.063	0.789	0.376	0.456	1.000

Notes: These are estimates of the true correlations between teachers' effects on each of the main outcomes. These estimates are based on the covariance matrix Ω^*. All measures are coded so that better teachers should improve the relevant outcome. (In particular, teacher effects on attendance is -1 * ln(Days Absent + 1).) In panel (A), Elementary school is defined as 5th grade and teachers teach both math and ELA. In panel (B), Middle school is defined as 6th-7th grade and test scores are for the subject the teacher teaches.

Table 4: Composition of Principal Components

	Component 1	Component 2	Component 3	Component 4
(A) Elementary School				
Math Test	0.393	0.617	0.383	-0.475
ELA Test	0.314	0.449	0.157	0.671
Attendance	-0.001	0.003	0.014	-0.001
Future Math Test	0.159	-0.103	-0.009	0.101
Future ELA Test	0.440	0.010	-0.591	-0.431
Future Attendance	0.436	-0.503	0.318	-0.038
Future Grades Math	0.386	0.046	-0.537	0.348
Future ELA Grades	0.435	-0.390	0.301	0.072
(B) Middle School				
Test Scores	0.144	0.625	-0.515	-0.127
Attendance	0.004	0.011	0.009	0.059
Current Grade	0.807	-0.328	-0.292	0.393
Future Test Scores	0.159	0.629	-0.029	0.138
Future Attendance	0.116	0.316	0.675	0.543
Future Grade	0.537	0.071	0.439	-0.715

Notes: This table reports the results of principal components analysis. The estimates indicate the composition of each of the first four components, estimated separately for elementary and middle school. Elementary school is defined as 5th grade and middle school is 6th-7th grade. For middle school, test scores and grades are for the one subject taught by the relevant teacher. Elementary school teachers teach both math and English, so additional outcomes are used for these teachers.

Table 5: Correlation of Estimates of Teacher Effectiveness

	Weighted Summary Measures			Empirical Bayes Estimates			
	PCA First Eigenvalue (1)	PCA Regression Coefficients (2)	Regression Coefficients (3)	Multi Dimension Test VA (4)	Single Dimension Test (5)	Multi Dimension Non-Test (6)	Single Dimension Non-Test (7)
(A) Elementary School							
PCA First Eigenvalue	1.000	0.929	0.944	0.698	0.510	0.821	0.615
PCA Regression	0.929	1.000	0.962	0.403	0.263	0.923	0.683
Regression	0.944	0.962	1.000	0.503	0.366	0.903	0.703
Multidim Test VA	0.698	0.403	0.503	1.000	0.617	0.297	0.159
Single Dim Test VA	0.510	0.263	0.366	0.617	1.000	0.133	0.183
Multidim Non-Test VA	0.821	0.923	0.903	0.297	0.133	1.000	0.550
Single Dim Non-Test VA	0.615	0.683	0.703	0.159	0.183	0.550	1.000
(B) Middle School							
PCA First Eigenvalue	1.000	0.946	0.847	0.246	0.177	0.950	0.798
PCA Regression	0.946	1.000	0.965	0.496	0.330	0.931	0.793
Regression	0.847	0.965	1.000	0.589	0.385	0.872	0.772
Multidim Test VA	0.246	0.496	0.589	1.000	0.485	0.244	0.171
Single Dim Test VA	0.177	0.330	0.385	0.485	1.000	0.102	0.199
Multidim Non-Test VA	0.950	0.931	0.872	0.244	0.102	1.000	0.715
Single Dim Non-Test VA	0.798	0.793	0.772	0.171	0.199	0.715	1.000

Notes: These estimates show the correlation between different measures of teacher effectiveness. The first three columns are based on the weighted summary measures of teacher effectiveness. Column 1 is based on the weights (coefficients) from a regression of teacher effects on high school graduation on the empirical Bayes estimates of teacher effects on individual outcomes. Column 2 is based on weights from a regression of teacher effects on high school graduation on the first four components from principal components analysis. Column 3 is based on weights from the first eigenvalue from principal components analysis. Column 4 is based on our estimate of teacher effects on test scores in the multidimensional setting. Column 5 is based on traditional estimates of teacher effects on test scores in the single dimension setting. Column 6 is based on our estimates of teacher effects on non-test score outcomes in the multidimensional setting. Column 7 is based on estimates of teacher effects on non-test outcomes in the single dimension setting. Non-test score empirical Bayes estimates are based on teacher effects on attendance, future attendance, future grades (and current grades for middle school). This measure equally weights teacher effects on these four outcomes. Panel (A) is based on elementary school teachers (grade 5) and panel (B) table is based on middle school teachers (grades 6-7). For elemenatry school, test VA is an average of the teacher's effect on math and ELA. For middle school, test VA is for the subject taught by the relevant teacher.

Table 6: Elementary School: Implications of Changing Evaluation Measures

	Weighted Summary Measures			Empirical Bayes Estimates	
	First Eigenvalue (1)	PCA Reg (2)	Reg (3)	Test Value-Added (4)	Non-Test Value-Added (5)
(A) Projected Change in Outcomes from Replacing Bottom 5% with Mean Teacher					
HS Graduation	0.149	0.145	0.142	0.078	0.111
Test Scores	0.524	0.203	0.230	0.907	0.207
Non-Test Outcomes	0.782	0.801	0.981	0.303	1.071
(B) Percent of Bottom 5% on Column VA also in Bottom 5% on Row VA					
HS Graduation	0.234	0.192	0.215	0.168	0.178
Test Scores	0.374	0.182	0.201	1.000	0.159
Non-Test Outcomes	0.467	0.481	0.696	0.159	1.000
Eigenvalue Summary	1.000	0.687	0.617	0.374	0.467

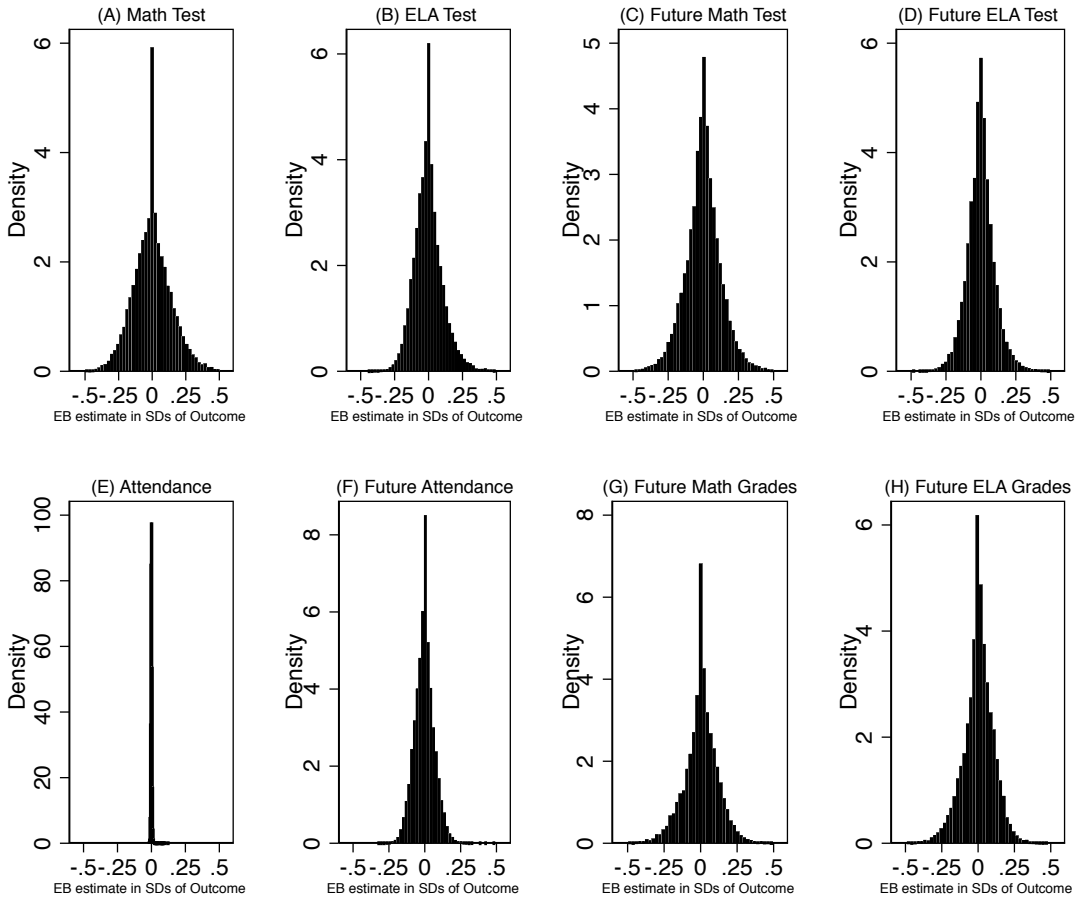
Notes: The estimates in Panel (A) show the differences in projected outcomes (high school graduation, test scores and non-test outcomes) for average teachers and those in the bottom 5% as ranked in terms of the value-added measure from the relevant column. The estimates in Panel (B) show the fraction of teachers in the bottom 5% in terms of the value-added metrics in the relevant row who are also in the bottom 5% in terms of the column's value-added metric. The first five columns are based on the weighted summary measures of teacher effectiveness. Column 1 is based on weights from the first eigenvalue from principal components analysis. Column 2 is based on weights from a regression of teacher effects on high school graduation on the first four components from principal components analysis. Column 3 is based on the weights (coefficients) from a regression of teacher effects on high school graduation on the empirical Bayes estimates of teacher effects on individual outcomes. Column 4 is based on traditional estimates of teacher effects on test scores in the single dimension setting. Column 5 is based on estimates of teacher effects on non-test outcomes in the single dimension setting. Non-test score empirical Bayes estimates are based on teacher effects on attendance, future attendance, future grades in subject and future grades in other subjects. This measure equally weights teacher effects on these four outcomes. This table is based on elementary school teachers (grade 5) and test score measurse are based on averages across math and reading.

Table 7: Middle School: Implications of Changing Evaluation Measures

	Weighted Summary Measures			Empirical Bayes Estimates	
	PCA First Eigenvalue (1)	PCA Regression Coefficients (2)	Regression Coefficients (3)	Test Value-Added (4)	Non-Test Value-Added (5)
(A) Projected Change in Outcomes from Replacing Bottom 5% with Mean Teacher					
HS Graduation	0.100	0.109	0.100	0.089	0.104
Test Scores	0.329	0.564	0.555	1.591	0.408
Non-Test Outcomes	0.993	1.028	1.030	0.214	1.147
(B) Percent of Bottom 5% on Column VA also in Bottom 5% on Row VA					
HS Graduation	0.138	0.151	0.146	0.215	0.182
Test Scores	0.137	0.206	0.207	1.000	0.176
Non-Test Outcomes	0.651	0.708	0.723	0.176	1.000
Eigenvalue Summary	1.000	0.757	0.604	0.137	0.651

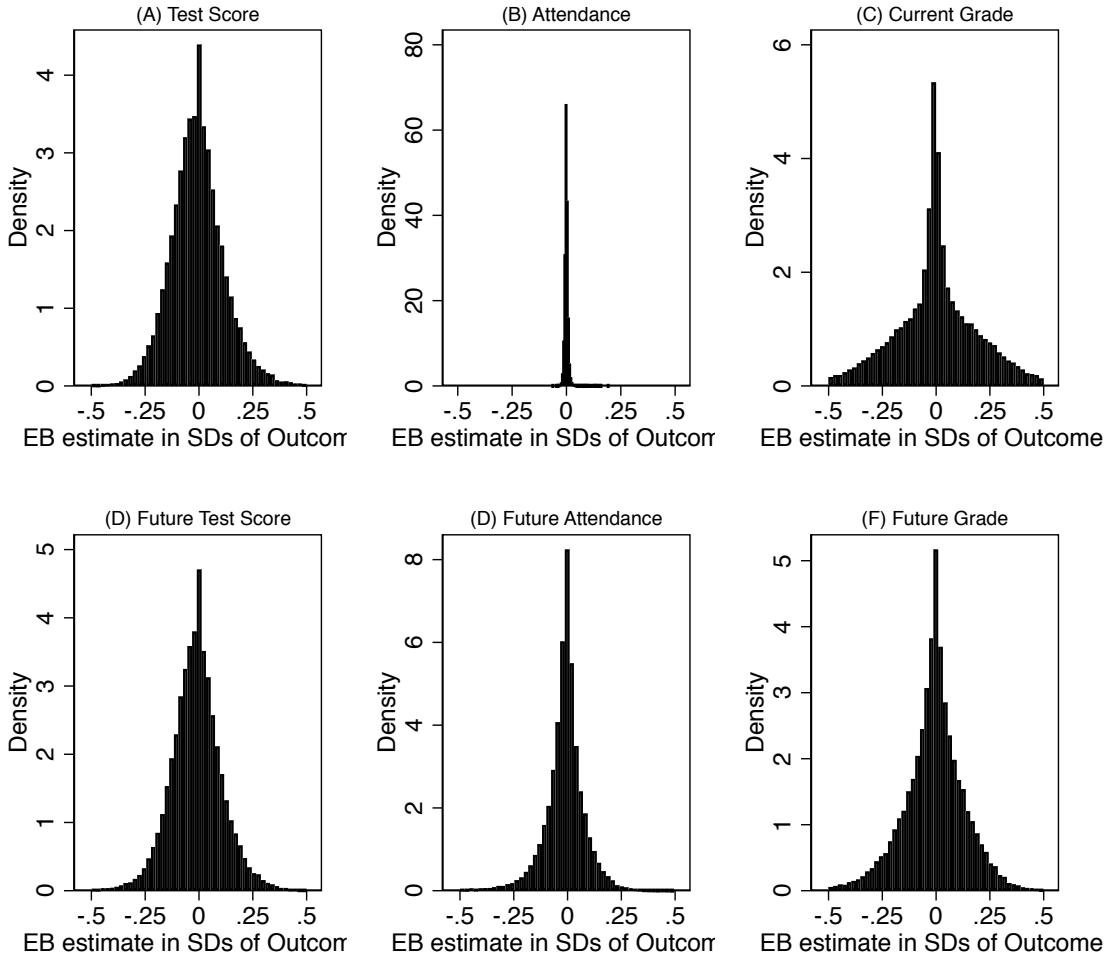
Notes: The estimates in Panel (A) show the differences in projected outcomes (high school graduation, test scores and non-test outcomes) for average teachers and those in the bottom 5% as ranked in terms of the value-added measure from the relevant column. The estimates in Panel (B) show the fraction of teachers in the bottom 5% in terms of the value-added metrics in the relevant row who are also in the bottom 5% in terms of the column's value-added metric. The first three columns are based on the weighted summary measures of teacher effectiveness. Column 1 is based on the weights (coefficients) from a regression of teacher effects on high school graduation on the empirical Bayes estimates of teacher effects on individual outcomes. Column 2 is based on weights from a regression of teacher effects on high school graduation on the first four components from principal components analysis. Column 3 is based on weights from the first eigenvalue from principal components analysis. Column 4 is based on traditional estimates of teacher effects on test scores in the single dimension setting. Column 5 is based on estimates of teacher effects on non-test outcomes in the single dimension setting. Non-test score empirical Bayes estimates are based on teacher effects on attendance, future attendance, current grades, and future grades. This measure equally weights teacher effects on these four outcomes. This table is based on middle school teachers (grades 6-7) and test score measurse are based on the subject a teacher teaches.

Figure 1: Distribution of Empirical Bayes Estimates for Elementary School Teachers



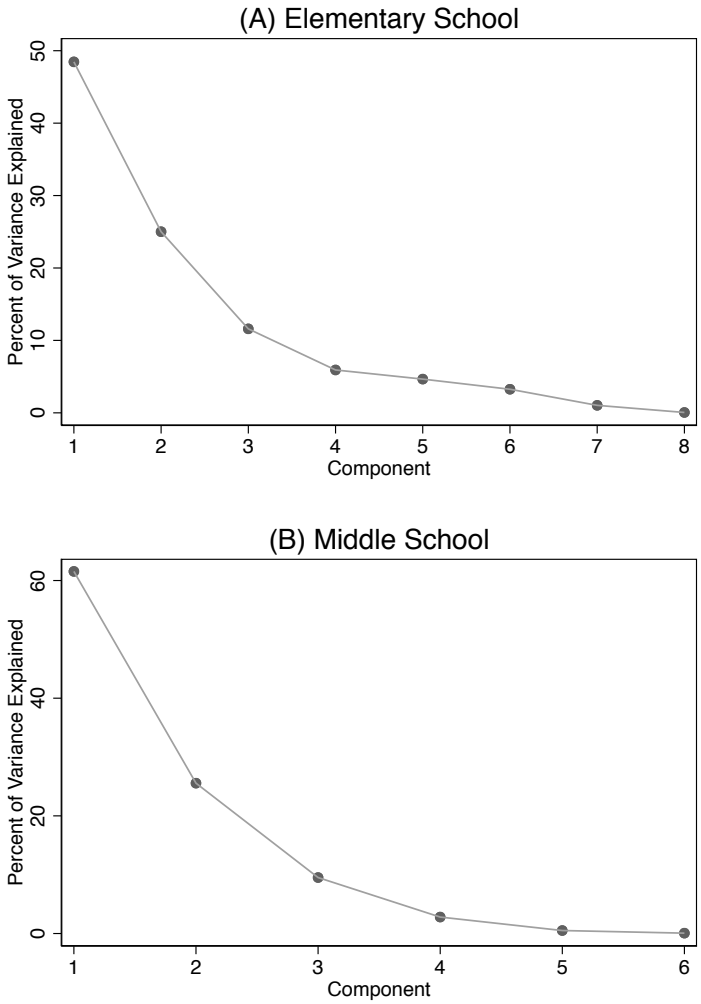
Notes: The figures above show the distribution of the multidimensional empirical Bayes estimates of teacher effects on individual outcomes for elementary school (5th grade). All estimates are in standard deviations of the outcome measure (standardized at the student level before computing teacher effects). Panels A, B and E are based on student outcomes in the year they are taught by the focal teacher. The remaining panels are based on student outcomes in the year following assignment to the focal teacher. Elementary school teachers teach both math and ELA.

Figure 2: Distribution of Empirical Bayes Estimates for Middle School Teachers



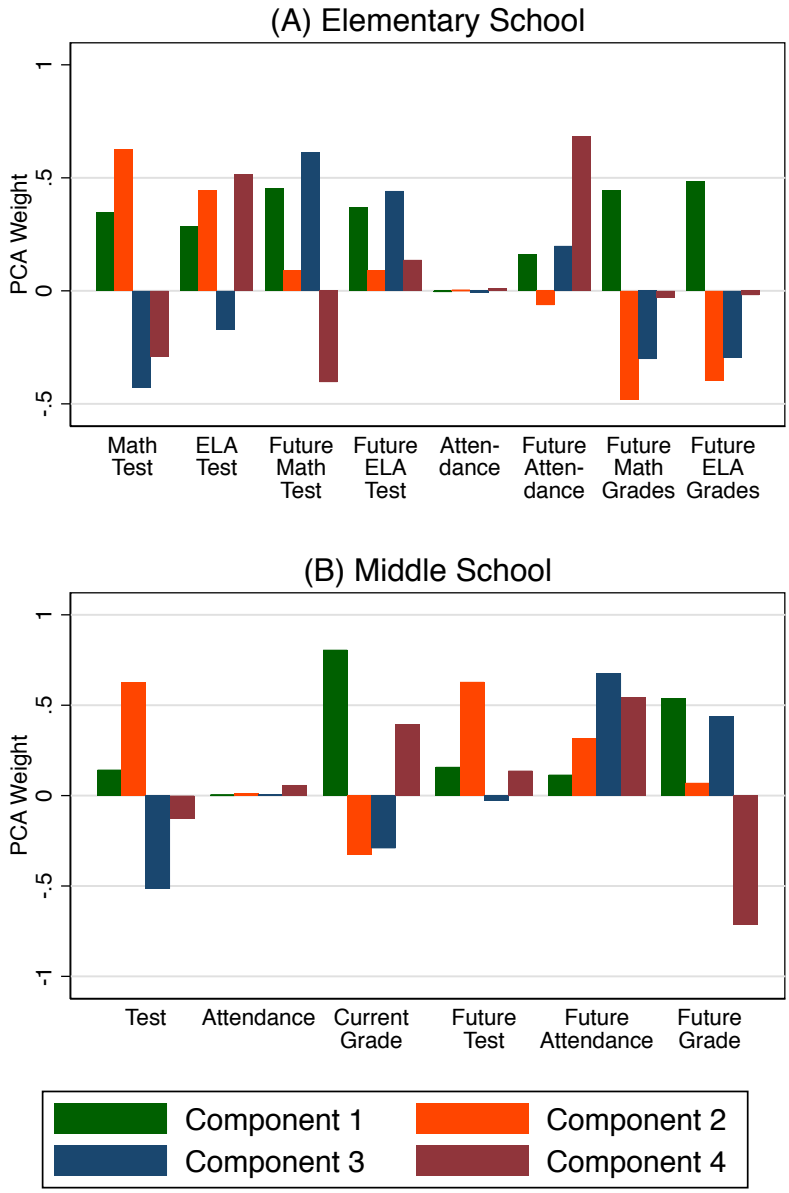
Notes: The figures above show the distribution of the multidimensional empirical Bayes estimates of teacher effects on individual outcomes for middle school (6th - 7th grade). All estimates are in standard deviations of the outcome measure (standardized at the student level before computing teacher effects). Panels A, B, and C are based on student outcomes in the year they are taught by the focal teacher. The remaining panels are based on student outcomes in the year following assignment to the focal teacher. Middle school teachers typically teach one grade so test score effects reflect the relevant subject.

Figure 3: Scree Plot of Eigenvalues



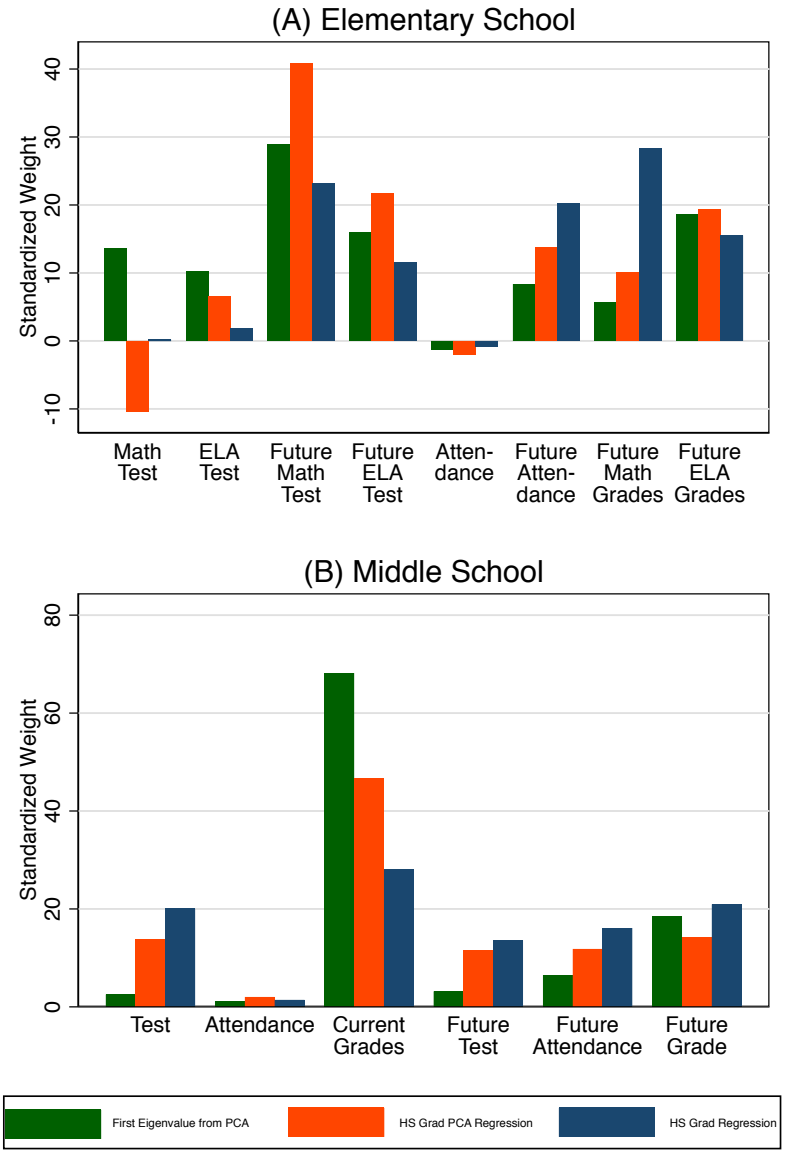
Notes: The figures above show the percent of variance in teacher effects on our student outcome measures explained by each principal component. These estimates come from conducting principal component analysis on the true measures of teacher effects. Panel (A) is for elementary school and is based on eight student outcome measures. Panel (B) is for middle school and is based on six student outcome measures.

Figure 4: PCA Components



Notes: The figures above show the relative weight each student outcome receives in each of the four main principal components. For middle school (in panel B) test scores and grades refer to the subject taught by the focal teacher. In elementary school (panel A) teachers teach both math and ELA. The principal components in panels A and B are not the same, in part because they are based on different sets of outcomes. For both middle and elementary school, the first four principal components each individually explain at least five percent of the variation in teacher effects on the relevant outcomes.

Figure 5: Composition of Weights



Notes: The figures above show the relative weights placed on each individual outcome for each of our three main approaches for creating summary measures of teacher effectiveness. The first approach (represented by the green bars) uses the first eigenvalue from principal component analysis to combine teacher effects on the outcomes into a summary measure. The height of the green bars shows the extent to which each individual outcome contributes to the summary measure. The second approach (orange bars) uses the coefficients from a regression of high school graduation on the four PCA components to weight individual outcomes in a summary measure. The third approach (navy bars) uses the coefficients from a regression of high school graduation on the empirical Bayes estimates of the individual outcomes as weights. Panel (A) shows the weights for elementary school teachers and panel (B) shows them for middle school teachers.

A Additional Tables and Figures

Table A.1: PCA: Proportion of Variance Explained by Components

	True Measures (1)	Empirical Bayes (2)	Raw Measures (3)
(A) Elementary School			
Component 1	0.487	0.618	0.423
Component 2	0.279	0.278	0.191
Component 3	0.113	0.053	0.105
Component 4	0.052	0.027	0.089
Component 5	0.045	0.014	0.075
Component 6	0.015	0.008	0.068
Component 7	0.008	0.001	0.046
Component 8	0.000	0.000	0.002
(B) Middle School			
Component 1	0.616	0.641	0.473
Component 2	0.256	0.278	0.200
Component 3	0.095	0.074	0.154
Component 4	0.028	0.006	0.113
Component 5	0.005	0.001	0.057
Component 6	0.001	0.000	0.002

Notes: These estimates indicate the proportion of variance explained by each component when conducting principal components analysis on the true measures of teacher effects (in column 1), the empirical Bayes measures (in column 2) and the raw measures of teacher effects. For elementary school, PCA is conducted on eight outcomes, and for middle school it is conducted on six outcomes.

Table A.2: Regression Results: Predictors of High School Graduation

	Elementary School		Middle School	
	Principal Components (1)	Individual Measures (2)	Principal Components (3)	Individual Measures (4)
(A) Principal Components				
First Component	0.029***		0.016***	
	(0.003)		(0.001)	
Second Component	-0.015***		0.010***	
	(0.004)		(0.001)	
Third Component	0.001		0.003***	
	(0.003)		(0.001)	
Fourth Component	0.000		-0.005***	
	(0.002)		(0.001)	
(B) Individual Measures				
Math Test		0.004		
		(0.006)		
ELA Test		0.007		
		(0.008)		
Test Score				0.016
				(0.010)
Attendance		-0.011*		-0.007***
		(0.007)		(0.002)
Current Grade in Subject				-0.001
				(0.005)
Future Math Test		-0.012		
		(0.016)		
Future ELA Test		-0.001		
		(0.005)		
Future Test Score				-0.009
				(0.014)
Future Attendance		0.041		0.009
		(0.030)		(0.008)
Future Grade Math		0.013		
		(0.010)		
Future Grade ELA		-0.021		
		(0.026)		
Future Grade in Subject				0.015**
				(0.006)
N	3,042	3,042	16,239	16,239

Notes: (* p<.10 ** p<.05 *** p<.01). Each observation is a teacher-subject-year. Columns 1 and 3 use the empirical Bayes estimates of the components that result from conducting PCA on the true measures of teacher effects. Columns 2 and 4 are based on the empirical Bayes estimates of effectiveness in terms of individual outcomes. Measures are standardized so that the coefficient represents the effect of a one standard deviation better teacher (in terms of that measure). The coefficients are from a regression of teacher effects on high school graduation for cohort $+1 on teacher effects on short-term outcomes for cohort$. We can only estimate teacher effects on high school graduation for 5th grade teachers in 2006 and 2007, and for middle school teachers in 2006-2010. Standard errors are clustered at the teacher-level.

Table A.3: Regression Results: Predictors of Regents Diplomas

	Elementary School		Middle School	
	Regents Diploma (1)	Advanced Regents (2)	Regents Diploma (3)	Advanced Regents (4)
(A) PCA Components				
First Component	0.031***	0.034***	0.015***	0.013***
	(0.003)	(0.003)	(0.001)	(0.001)
Second Component	-0.015***	-0.013***	0.010***	0.016***
	(0.004)	(0.003)	(0.001)	(0.001)
Third Component	0.002	0.001	0.003***	-0.005***
	(0.003)	(0.003)	(0.001)	(0.001)
Fourth Component	-0.002	-0.005**	-0.005***	-0.001
	(0.002)	(0.002)	(0.001)	(0.001)
(B) Individual Measures				
Math Test Score	0.005	0.005		
	(0.006)	(0.006)		
ELA Test Score	0.007	0.011		
	(0.008)	(0.007)		
Test Score			0.023**	0.010
			(0.010)	(0.007)
Attendance	-0.012*	-0.015**	-0.009***	-0.001
	(0.006)	(0.006)	(0.002)	(0.002)
Current Grade in Subject			0.000**	0.004
			(0.005)	(0.004)
Future Math Test	- 0.010	- 0.007		
	(0.015)	(0.013)		
Future ELA Test	0.003	0.007		
	(0.005)	(0.005)		
Future Test Score			-0.019	0.010
			(0.013)	(0.010)
Future Attendance	0.039	0.052**	0.015*	-0.002
	(0.030)	(0.023)	(0.008)	(0.005)
Future Grade in Subject			0.012*	0.004
			(0.006)	(0.004)
Future Math Grade	0.005	0.003		
	(0.010)	(0.006)		
Future ELA Grade	-0.018	-0.034*		
	(0.026)	(0.019)		
N	3,042	3,042	16,239	16,239

Notes: (* p<.10 ** p<.05 *** p<.01). Each observation is a teacher-subject-year. Panel (A) uses the empirical Bayes estimates of the components that result from conducting PCA on the true measures of teacher effects. Panel (B) is based on the empirical Bayes estimates of effectiveness in terms of individual outcomes. Measures are standardized so that the coefficient represents the effect of a one standard deviation better teacher (in terms of that measure). The coefficients are from a regression of teacher effects on high school graduation for cohort $+1 on teacher effects on short-term outcomes for cohort$. We can only estimate teacher effects on high school graduation and regents diplomasfor 5th grade teachers in 2006 and 2007, and for middle school teachers in 2006-2010. Standard errors are clustered at the teacher-level.

Table A.4: Composition of Weights

	Unstandardized Weights			Standardized Weights		
	First Eigenvalue (1)	PCA Regression (2)	Regression (3)	First Eigenvalue (4)	PCA Regression (5)	Regression (6)
(A) Elementary School						
Math Test	14.290	-11.670	0.142	13.634	-10.332	0.151
ELA Test	13.618	9.244	2.154	10.276	6.473	1.808
Attendance	33.044	50.465	23.878	28.835	40.864	23.176
Future Math Test	22.429	33.027	14.717	15.905	21.732	11.608
Future ELA Test	-24.247	-41.447	-15.458	-1.237	-1.963	-0.878
Future Attendance	14.737	26.172	32.063	8.367	13.789	20.249
Future Math Grade	6.114	11.733	27.438	5.682	10.118	28.362
Future ELA Grade	20.015	22.476	15.067	18.538	19.318	15.523
(B) Middle School						
Test Scores	3.105	13.257	18.959	2.523	13.824	20.030
Attendance	16.690	24.112	16.788	1.090	2.021	1.426
Current Grade	47.259	25.169	14.917	68.178	46.601	27.983
Future Test Scores	4.201	11.973	13.799	3.165	11.579	13.521
Future Attendance	10.189	14.469	19.432	6.494	11.836	16.105
Future Grade	18.557	11.020	16.105	18.550	14.138	20.935

Notes: This table shows how much each individual component is weighted in our three main weighting approaches. Each observation is a teacher-subject-year. Columns 1 and 4 contain weights based on the first eigenvalue from principal components analysis. Columns 2 and 5 contain weights based on the coefficients from a regression of teacher effects on high school graduation on the first four components from principal components analysis. Columns 3 and 6 contains weights based on the coefficients from a regression of teacher effects on high school graduation on the empirical Bayes estimates of teacher effects on individual outcomes. The weights in columns 4 to 6 are standardized to account for the variation in teacher effects on each of the outcomes. The weights for elementary school (5th grade) are in panel A and those for middle school (6th-7th grade) are in panel B.

Table A.5: Weights based on Predicting Regents Diploma and Advanced Regents

	Unstandardized Weights				Standardized Weights			
	Regents Diploma		Advanced Regents		Regents Diploma		Advanced Regents	
	PCA Regression (1)	Regression Coefficients (2)	PCA Regression (3)	Regression Coefficients (4)	PCA Regression (5)	Regression Coefficients (6)	PCA Regression (7)	Regression Coefficients (8)
(A) Elementary School								
Math Test	-8.159	4.994	-2.414	10.946	-7.291	5.326	-2.201	11.281
ELA Test	10.042	1.322	11.508	0.314	7.098	1.115	8.297	0.256
Attendance	52.099	26.899	55.594	37.070	42.581	26.235	46.348	34.942
Future Math Test	32.519	15.808	31.887	22.697	21.599	12.529	21.603	17.386
Future ELA Test	-39.902	-14.782	-37.036	-19.402	-1.907	-0.843	-1.806	-1.070
Future Attendance	25.100	31.198	23.491	29.048	13.348	19.798	12.742	17.816
Future Math Grade	7.550	18.041	-0.327	15.795	6.571	18.739	-0.291	15.856
Future ELA Grade	20.751	16.520	17.298	3.532	18.002	17.102	15.307	3.534
(B) Middle School								
Test Scores	13.707	19.518	25.643	37.065	14.404	22.324	29.826	38.399
Attendance	24.128	23.940	24.077	13.199	2.038	2.201	2.251	1.099
Current Grade	24.372	14.295	13.978	14.101	45.477	29.033	28.868	25.938
Future Test Scores	12.310	10.604	16.472	19.571	11.998	11.249	17.769	18.804
Future Attendance	14.652	18.309	13.672	10.212	12.078	16.429	12.475	8.299
Future Grade	10.832	13.334	6.158	5.853	14.005	18.764	8.812	7.461

Notes: This table shows the weights from the PCA regression and regression approach when use Regents Diploma receipt or Advanced Regents Diploma as the long-term outcome of interest. Each observation is a teacher-subject-year. Columns 1 and 5 contain weights based on the coefficients from a regression of teacher effects on Regents diploma receipt on the first four components from principal components analysis. Columns 2 and 6 contain weights based on the coefficients from a regression of teacher effects on Regents diploma receipt on the empirical Bayes estimates of teacher effects on individual outcomes. Columns 3 and 7 contain weights based on the coefficients from a regression of teacher effects on earning an Advanced Regents diploma receipt on the first four components from principal components analysis. Columns 4 and 8 contain weights based on the coefficients from a regression of teacher effects on earning an Advanced Regents diploma on the empirical Bayes estimates of teacher effects on individual outcomes. The weights in columns 5 through 8 are standardized to account for the variation in teacher effects on each of the eight outcomes.

Table A.6: Correlation between Multidimensional and Single Dimension Empirical Bayes Estimates

(A) Elementary School	Math Test Scores	ELA Test Scores	Attendance	Future Math Test	Future ELA Test	Future Attendance	Future Math Grades	Future ELA Grades
Correlation	0.597	0.326	0.113	0.287	0.419	0.333	0.441	0.507

(B) Middle School	Test Scores	Attendance	Grade in Subject	Future Test Scores	Future Attendance	Future Grade in Subject
Correlation	0.365	0.088	0.766	0.444	0.405	0.656

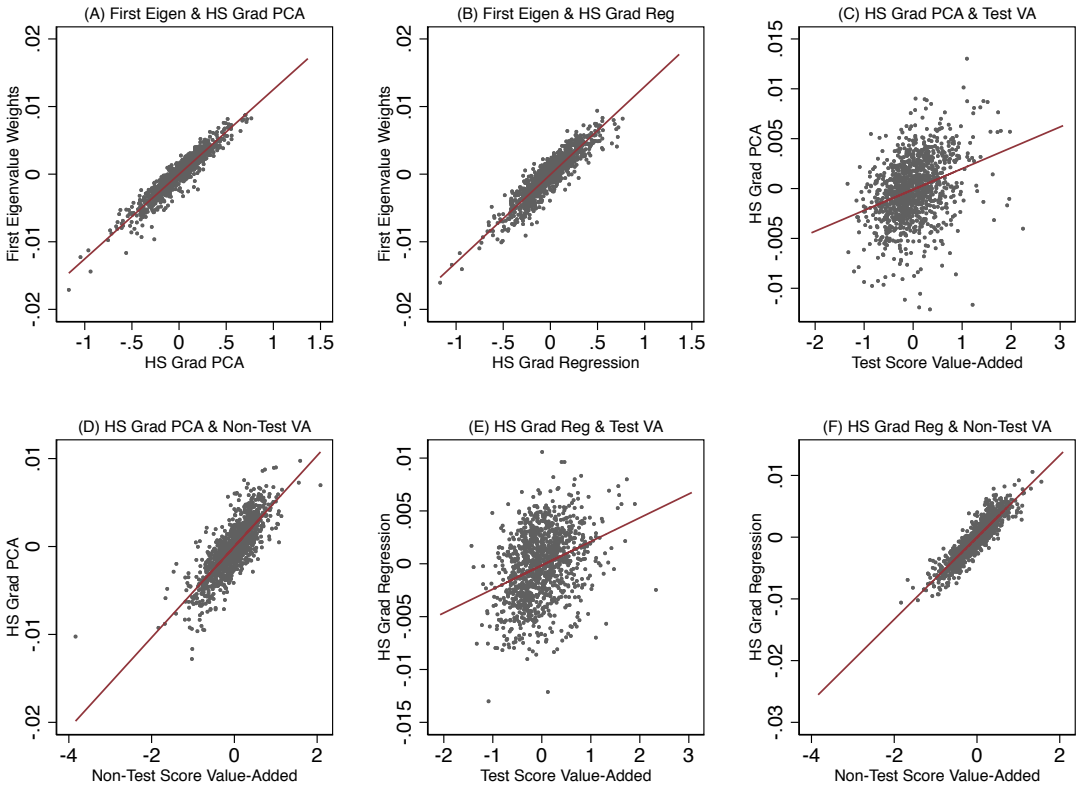
Notes: Panel A is based on elementary school (5th grade) teachers and panel B is based on middle school teachers (6th-7th grade). Estimates indicate the correlation between the single and multidimensional empirical Bayes' estimates of teacher effects on the noted outcome. The multidimensional empirical Bayes estimates incorporate information about teacher effects on and the noisiness of other outcomes.

Table A.7: Spearman Correlations of Estimates of Teacher Effectiveness

	Weighted Summary Measures			Empirical Bayes Estimates			
	PCA First Eigenvalue (1)	PCA Regression Coefficients (2)	Regression Coefficients (3)	Multi Dimension Test VA (4)	Single Dimension Test (5)	Multi Dimension Non-Test (6)	Single Dimension Non-Test (7)
(A) Elementary School							
PCA First Eigenvalue	1.000	0.922	0.940	0.701	0.625	0.838	0.782
PCA Regression	0.922	1.000	0.960	0.409	0.336	0.955	0.873
Regression	0.940	0.960	1.000	0.504	0.454	0.933	0.890
Multidim Test VA	0.701	0.409	0.504	1.000	0.941	0.292	0.274
Single Dim Test VA	0.625	0.336	0.454	0.941	1.000	0.218	0.206
Multidim Non-Test VA	0.838	0.955	0.933	0.292	0.218	1.000	0.933
Single Dim Non-Test VA	0.782	0.873	0.890	0.274	0.206	0.933	1.000
(B) Middle School							
PCA First Eigenvalue	1.000	0.946	0.849	0.255	0.272	0.967	0.934
PCA Regression	0.946	1.000	0.964	0.490	0.468	0.945	0.923
Regression	0.849	0.964	1.000	0.586	0.550	0.883	0.896
Multidim Test VA	0.255	0.490	0.586	1.000	0.931	0.244	0.236
Single Dim Test VA	0.272	0.468	0.550	0.931	1.000	0.222	0.238
Multidim Non-Test VA	0.967	0.945	0.883	0.244	0.222	1.000	0.961
Single Dim Non-Test VA	0.934	0.923	0.896	0.236	0.238	0.961	1.000

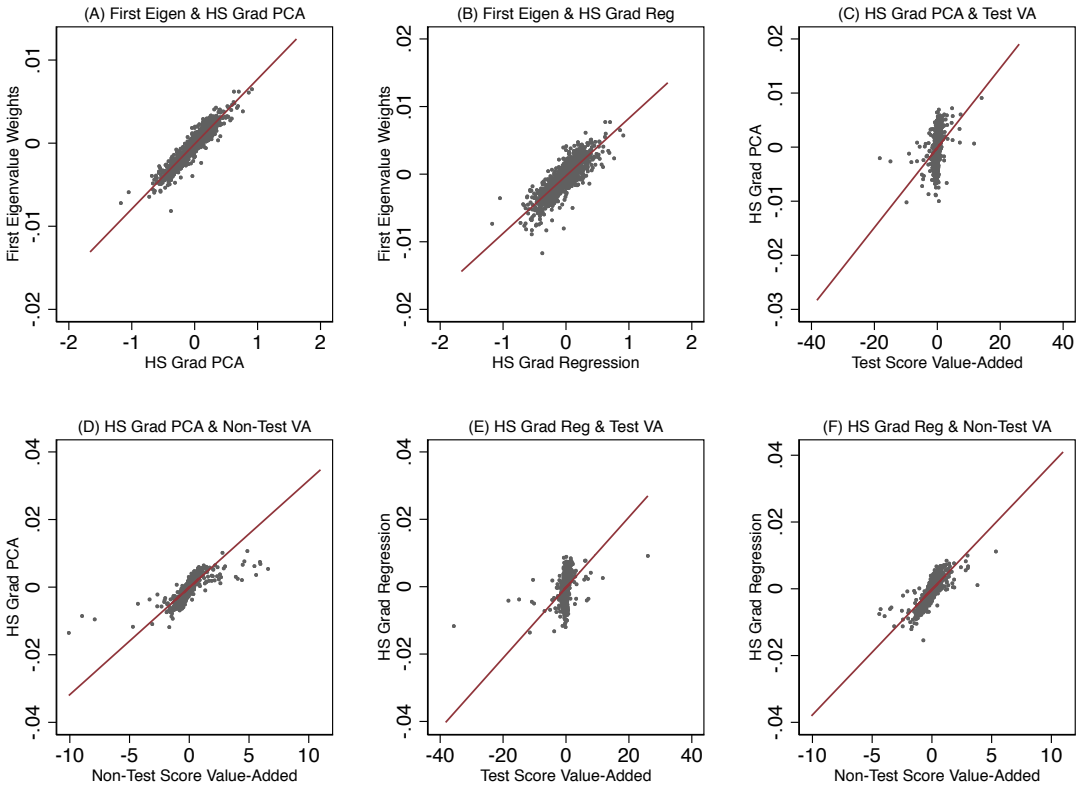
Notes: These estimates show the Spearman rank correlations between different measures of teacher effectiveness. (This is a non-parametric estimate of the association between two measures.) The first three columns are based on the weighted summary measures of teacher effectiveness. Column 1 is based on the weights (coefficients) from a regression of teacher effects on high school graduation on the empirical Bayes estimates of teacher effects on individual outcomes. Column 2 is based on weights from a regression of teacher effects on high school graduation on the first four components from principal components analysis. Column 3 is based on weights from the first eigenvalue from principal components analysis. Column 4 is based on our estimate of teacher effects on test scores in the multidimensional setting. Column 5 is based on traditional estimates of teacher effects on test scores in the single dimension setting. Column 6 is based on our estimates of teacher effects on non-test score outcomes in the multidimensional setting. Column 7 is based on estimates of teacher effects on non-test outcomes in the single dimension setting. Non-test score empirical Bayes estimates are based on teacher effects on attendance, future attendance, future grades, (and current grades for middle school). This measure equally weights teacher effects on these four outcomes. Panel (A) is based on elementary school teachers (grade 5) and panel (B) table is based on middle school teachers (grades 6-7). For elemenatry school, test VA is an average of the teacher's effect on math and ELA. For middle school, test VA is for the subject taught by the relevant teacher.

Figure A.1: Elementary School: Correlation between Teacher Ratings on Different Measures of Effectiveness



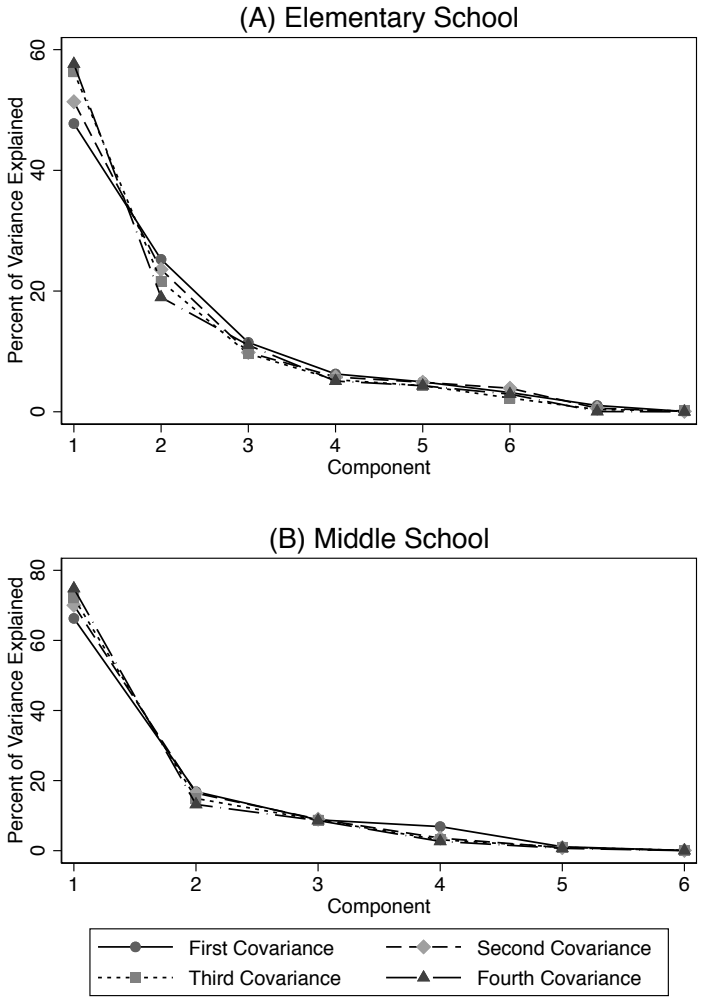
Notes: These figures show the correlations between elementary school teachers' ratings on our summary measures of effectiveness and traditional single dimensions value-added measures based on test scores or non-test score measures. In panels (A) and (B), the y-axis is based on the weights from the *first eigenvalue* from principal component analysis. The y-axis in panels (C) and (D) is based on the (*PCA regression*) approach which uses the weights from a regression of high school graduation on the four PCA components. In panels (E) and (F), the y-axis shows the summary measure based on a regression of high school graduation on the empirical Bayes estimates of the individual outcomes. Panels (C) and (E) look at correlations with test score value-added and Panels (D) and (F) look at correlations with non-test value-added. The dots represent the standardized ratings for individual teachers and the red lines show the relationship between the two relevant measures. These figures are for elementary school teachers who teach fifth grade.

Figure A.2: Middle School: Correlation between Teacher Ratings on Different Measures of Effectiveness



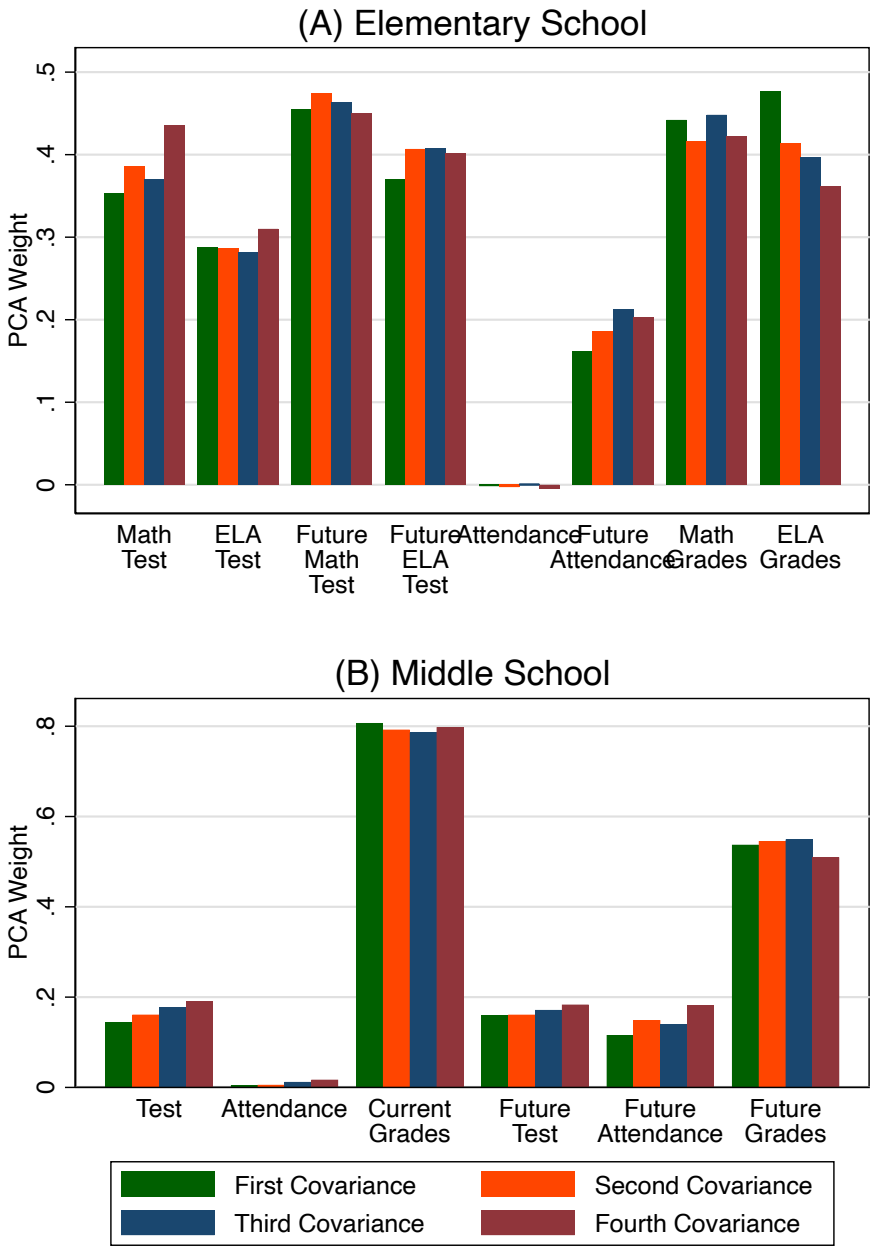
Notes: These figures show the correlations between middle school teachers' ratings on our summary measures of effectiveness and traditional single dimensions value-added measures based on test scores or non-test score measures. In panels (A) and (B), the y-axis is based on the weights from the *first eigenvalue* from principal component analysis. The y-axis in panels (C) and (D) is based on the (*PCA regression*) approach which uses the weights from a regression of high school graduation on the four PCA components. In panels (E) and (F), the y-axis shows the summary measure based on a regression of high school graduation on the empirical Bayes estimates of the individual outcomes. Panels (C) and (E) look at correlations with test score value-added and Panels (D) and (F) look at correlations with non-test value-added. The dots represent the standardized ratings for individual teachers and the red lines show the relationship between the two relevant measures. These figures are for middle school teachers who teach sixth and seventh.

Figure A.3: Scree Plot of Eigenvalues



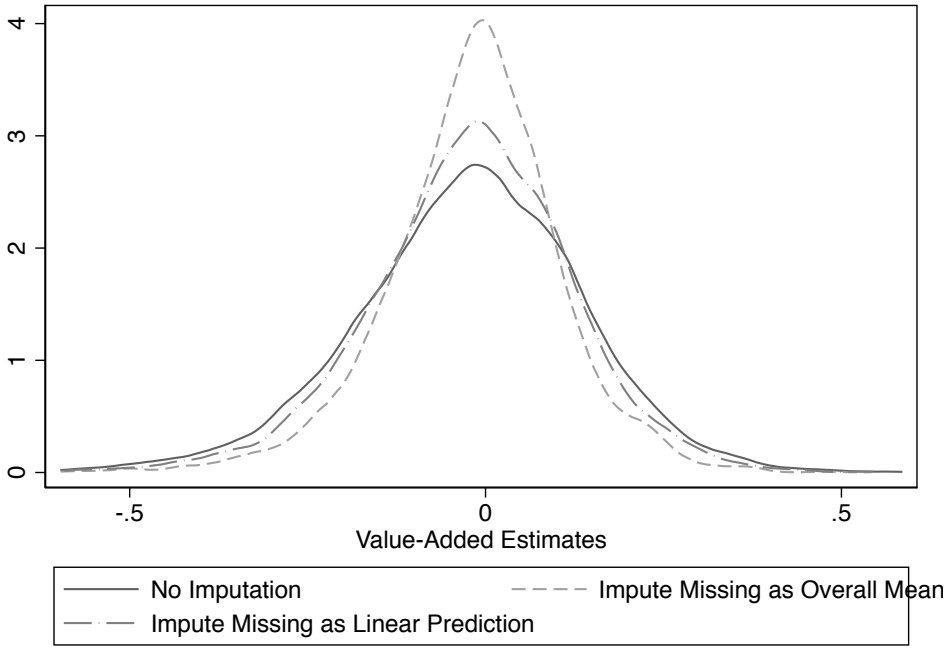
Notes: The figures above show the percent of variance in teacher effects on our student outcome measures explained by each principal component. These estimates come from conducting principal component analysis on four covariance matrices: $\mathbb{C}_1$, $\mathbb{C}_2$, $\mathbb{C}_3$, $\mathbb{C}_4$, where $\mathbb{C}_k$ is the covariance between $\Theta_{j,t}$ and $\Theta_{j,t-k}$. Panel (A) is for elementary school and is based on eight student outcome measures. Panel (B) is for middle school and is based on six student outcome measures.

Figure A.4: PCA Components



Notes: The figures above show the relative weight each student outcome receives in each of the first main principal components. For middle school (in panel B) test scores and future grades refer to the subject taught by the focal teacher. In elementary school (panel A) teachers teach both math and ELA. The principal components in panels A and B are not the same, in part because they are based on different sets of outcomes. For both middle and elementary school, we show the first principal component for each of the following four covariance matrices: $\mathbb{C}_1$, $\mathbb{C}_2$, $\mathbb{C}_3$, and $\mathbb{C}_4$, where $\mathbb{C}_k$ is the covariance between $\Theta_{j,t}$ and $\Theta_{j,t-k}$.

Figure A.5: Comparing the Missing Data Approaches



Notes: The figure above shows three empirical Bayes distributions. All three are a weighted average of the empirical Bayes estimates of all the effectiveness measures, with the weights estimated using the PCA Regression approach defined in Section II. They differ in how the missing observations are handled. The "No Imputation" method uses the approach defined in Section D. The other two approaches imputes the missing data and then estimates the empirical Bayes estimates as if none of the observations were missing. The "Impute Missing as Overall Mean" imputes the missing data at the overall mean and the "Impute Missing as Linear Prediction" imputes the missing observations as the best linear predictions conditional on the observed outcomes.

B Framework with Teacher Value-Added Drift

B.1 Model with Drift

In our main analysis, we assumed that teachers do not get more or less effective over time; instead, any teacher's effect on their students' outcomes is a combination of the teacher's persistent effectiveness and a year-specific shock. We now present the model in which there is drift and discuss how that changes the interpretation of the results; the model also informs the discussion in Section **??** of how to include multiple years of teacher effectiveness in the predictions.

As before, we can write the statistical model of student outcomes as:

$$y_{i,t-1} = \beta X_{i,t-1} + \Theta_{j,t-1} + \nu_{j,t-1} + \epsilon_{i,t-1} \tag{19}$$

where $X_{i,t-1}$ are the student's characteristics, $\Theta_{j,t-1}$ is the effect of the teacher on her outcomes, and both $\nu_{j,t-1}$ and $\epsilon_{i,t-1}$ are normally distributed error terms that represent the classroom and individual-shock, respectively. Note that we are slightly abusing notation here, in that before $\nu_{j,t-1}$ denoted the classroom shocks caused by both idiosyncratic shocks to the teachers' effectiveness and classroom shocks that have other causes and here $\nu_{j,t-1}$ only corresponds to classroom shocks caused by factors other than the teachers' effectiveness.

Defining a teacher's value-added in year $t-1$ as we do in Equation (4) and continuing to denote these estimates $\theta_{j,t-1}$, from this statistical model we get that if $\hat{\beta} \to \beta$:

$$\theta_{j,t-1}|\Theta_{j,t-1} \sim N\Big(\Theta_{j,t-1}, \Sigma_\nu + \frac{1}{N_j}\Sigma_\epsilon\Big) \tag{20}$$

If we assume that $\Theta_{j,t-1} \sim N(0, \Omega)$, we can then use Bayes' Law as before to show that:

$$\Theta_{j,t-1}|\theta_{j,t-1} \sim N\Big(\Omega_j^* \theta_{j,t-1}, \Sigma_j^*\Big) \tag{21}$$

where again

$$\Omega_j^* = (\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}$$
$$\Sigma_j^* = (\Sigma_j^{-1} + \Omega^{-1})^{-1}$$

The challenge is that we do not want the posterior distribution of $\Theta_{j,t-1}$ conditional on $\theta_{j,t-1}$ and instead want the posterior of $\Theta_{j,t}$ conditional on $\theta_{j,t-1}$. To calculate this posterior, we need to augment that model by specifying how $\Theta_{j,t-1}$ is linked to $\Theta_{j,t}$.

In the Section II.A, we linked $\Theta_{j,t-1}$ and $\Theta_{j,t}$ by assuming that both are equal to some

permanent component of teacher effectiveness and a year specific shock. We now relax that assumption and only assume that $\Theta_{j,t}$ evolves in a stationary Gaussian process. That is, we assume that:

$$\begin{bmatrix} \Theta_{j,t} \\ \Theta_{j,t-1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega & \mathbb{C}_1 \\ \mathbb{C}_1 & \Omega \end{bmatrix} \right) \tag{22}$$

for every t. Here Ω equals the variance of $\Theta_{j,t}$, while before we used it to only denote the persistent component of teacher effectiveness, and $\mathbb{C}_1$ equals the covariance of $\Theta_{j,t}$ and $\Theta_{j,t-1}$. Our assumption that they evolve in a Gaussian process implies that their joint distribution is distributed according to a multivariate normal distribution and the assumption that it's stationary implies that the variance and covariance of this distribution do not depend on t.

From this model, we get that:

$$\Theta_{j,t}|\Theta_{j,t-1} \sim N\Big(\mathbb{C}_1\Omega^{-1}\Theta_{j,t-1}, \Omega - \mathbb{C}_1\Omega^{-1}\mathbb{C}_1\Big) \tag{23}$$

which follows from the fact that conditioning on a portion of the observations in a multivariate normal distribution still results in a multivariate normal distribution. Our earlier assumption that $\Theta_{j,t}$ consists of a persistent component and year-specific shock, which we ignored when computing Ω, meant that $\mathbb{C}_1 = \Omega$. That meant that $\Theta_{j,t}|\Theta_{j,t-1} \sim N\big(\Theta_j, \mathbf{0}\big)$, which is why we were able to ignore this step in the derivations used in the main body of the paper.

We can then combine Equations (23) and (21) to get that:[33]

$$\Theta_{j,t}|\theta_{j,t-1} \sim N\Big(\mathbb{C}_1\Omega^{-1}\Omega_j^*\theta_{j,t-1}, \Omega - \mathbb{C}_1\Omega^{-1}\mathbb{C}_1 + \mathbb{C}_1\Omega^{-1}\Sigma_j^*\Big) \tag{24}$$

Substituting in the fact that $\Omega_j^* = \Omega(\Omega+\Sigma_j)^{-1}$, we get that the empirical Bayes estimates in a model with drift are the mean posterior, or:

$$\hat{\Theta}_{j,t} = \mathbb{C}_1(\Omega + \Sigma_j)^{-1}\theta_{j,t-1} \tag{25}$$

[33]To see why this is true, we can write $\Theta_{j,t} = \mathbb{C}_1\Omega^{-1}\Theta_{j,t-1} + \epsilon$ and $\Theta_{j,t-1} = \Omega_j^*\theta_{j,t-1} + \eta$, where ϵ and η are mean-zero and normally distributed error terms; these error terms are not to be confused with the ϵ and η error terms defined in the paper above and are instead placeholders for the error terms implied by the distributions of $\Theta_{j,t}|\Theta_{j,t-1}$ and $\Theta_{j,t-1}|\theta_{j,t-1}$. We can then combine these equations to get that $\Theta_{j,t} = \mathbb{C}_1\Omega^{-1}\Omega_j^*\theta_{j,t-1} + \mathbb{C}_1\Omega^{-1}\eta + \epsilon$. Note that $\theta_{j,t-1}$ is independent from ϵ since $\Theta_{j,t}|\Theta_{j,t-1}, \theta_{j,t-1} = \Theta_{j,t}|\Theta_{j,t-1}$. We therefore get that $\Theta_{j,t}|\theta_{j,t-1}$ is distributed normally with mean $\mathbb{C}_1\Omega^{-1}\Omega_j^*\theta_{j,t-1}$ and variance defined by the variance of $\mathbb{C}_1\Omega^{-1}\eta + \epsilon$.

B.2 Interpretation of Estimates

In the paper, we framed the results in a model without drift. We now briefly discuss how the interpretation changes in our model with drift in teacher effectiveness. To ease the comparison, we will put a bit more structure on the nature of the drift and assume that the teacher effects can be decomposed into three components: a persistent effect Θ_j, a time-varying effect denoted $\phi_{j,t}$, and a year-specific shock $\eta_{j,t}$. We will further assume that the time-varying component evolves according to a stationary AR(1) process, so $\phi_{j,t} = \rho\phi_{j,t-1} + \tilde{\phi}_{j,t}$ for $\rho \in (0,1)$ and an idiosyncratic error term $\tilde{\phi}_{j,t}$. Assuming the three components are independent, we then get that $\Omega = \mathbb{V}\big(\Theta_j\big) + \mathbb{V}\big(\phi_{j,t}\big) + \mathbb{V}(\eta_{j,t})$ and $\mathbb{C}_1 = \mathbb{V}\big(\Theta_j\big) + \rho\mathbb{V}\big(\phi_{j,t}\big)$

Our results regarding the dimensionality of teacher effectiveness therefore incorrectly attempted to explain how well $\Theta_j + \rho\phi_{j,t-1}$ could be summarized by a lower dimensional vector rather than $\Theta_j + \phi_{j,t-1}$. However, these differ only by $(1-\rho)\mathbb{V}\big(\phi_{j,t}\big)$. Unless most of the variation in teacher effectiveness is generated by the time-varying component (i.e. $\mathbb{V}\big(\phi_{j,t}\big)$ is much bigger than $\mathbb{V}(\Theta_j)$, ρ is much smaller than one, and the variance structure of Θ_j is quite different than the variance structure of $\phi_{j,t-1}$ the results are likely to be similar.

Furthermore there is also a conceptual justification for using the model without drift for our purposes. Fundamentally, $\Theta_j + \rho\phi_{j,t-1}$ is the only part of the teachers' effectiveness in year t that is knowable in year $t-1$. Just as we ignored the year-specific shock $\eta_{j,t}$ when exploring how well teacher effectiveness can be explained by a lower dimensional vector, one could argue that we should only be concerned with how well $\Theta_j + \rho\phi_{j,t-1}$ can be summarized rather than $\Theta_j + \phi_{j,t-1}$. From that perspective, it is actually $\mathbb{C}_1$ that we want to explain, rather than Ω, and the approach we use in the paper provides the correct empirical estimates, albeit motivated in a slightly incorrect way.

Having said that, we can also provide some empirical evidence that our results, which aim to understand the dimensionality of $\Theta_j + \rho\phi_{j,t-1}$, provide a similar results as if we were to explore the dimensionality of $\Theta_j + \phi_{j,t-1}$. While we cannot test directly how much our results would change if we used $\Theta_j + \phi_{j,t-1}$ instead of $\Theta_j + \rho\phi_{j,t-1}$ without more assumptions to better separate the classroom shock due to the teacher from the classroom shock not due to the teacher, we can explore whether our results change when looking at $\Theta_j + \rho^2\phi_{j,t-1}$ rather than $\Theta_j + \rho\phi_{j,t-1}$ by conducting a PCA on $\hat{\mathbb{C}}_2 = Cov(\theta_{i,t}, \theta_{i,t-2})$, rather than on $\hat{\mathbb{C}}_1$. If the results are similar, then it's likely that they would also be similar when exploring $\Theta_j + \phi_{j,t-1}$.

In Figures A.3 and A.4, we show that the results of the PCA are nearly identical, regardless of whether we estimate the components using $\mathbb{C}_1$, $\mathbb{C}_2$, $\mathbb{C}_3$, or $\mathbb{C}_4$. More specifically, in Figure A.3 we illustrate that the components explain a similar percentage of the overall variance regardless of the lag we use. In Figure A.4, we further show that the weights derived

from the first component are similar regardless of the lag used. We therefore believe that the results do not depend on the fact that we assumed $\mathbb{C}_1 \approx \Omega$.

Finally, although we used the model without drift to compute the empirical Bayes' estimates, the empirical Bayes estimates will actually be identical to those computed in a model with drift. Interestingly, this is true in spite of the fact that erroneously assuming away drift implies leads us to estimate both Ω and Σ_j incorrectly. However, since we estimate Σ_ν as consisting of the the "unexplained" variance of $\theta_{j,t}$, which corresponds to the unexplained variance of $\Omega + \Sigma_j$, we still correctly estimate $\Omega + \Sigma_j$ even though both the estimates of Ω and Σ_j are incorrect. For the empirical Bayes' estimates, therefore, specifying whether there is drift in teacher effectiveness or not is only important when including multiple years of data in the estimates; we discuss this more below.

C Empirical Bayes Estimates as Best Linear Predictors

The empirical Bayes estimates are equivalent to the best linear predictors of the true teacher effects given the previous years' estimated teacher effects. Formally, suppose we aim to know what weights Ψ_j^{k*} minimize the mean-squared error of the predicted teacher effect on measure K given $\theta_{j,t-1}$, or:[34]

$$\Psi_j^{k*} = \arg\min_{\Psi^k} \mathbb{E}\Big[\big(\Theta_j^k - \Psi^k\theta_{j,t-1}\big)'\big(\Theta_j^k - \Psi^k\theta_{j,t-1}\big)\Big] \tag{26}$$

It is clear from this specification, that Ψ_j^{k*} are just the coefficients from an OLS regression of Θ_j^k on $\theta_{j,t-1}$. Thus, we get that:

$$\Psi_j^{k*} = \mathbb{E}\Big[\Big((\theta'_{j,t-1}\theta_{j,t-1})^{-1}\theta'_{j,t-1}\Theta_j^k\Big)'\Big] \tag{27}$$

which implies that $\Psi_j^{k*} = \Big((\Omega+\Sigma_j)^{-1}\Omega^k\Big)'$, where Ω^k is the k^{th} column of covariance matrix of Θ_j. Combining the K estimates of Ψ_j^{k*}, we get that $\Phi^* = \big((\Omega + \Sigma_j)^{-1}\Omega\big)'$. Although this expression appears different, it turns out that $(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1} = \big((\Omega + \Sigma_j)^{-1}\Omega\big)'$. Thus, the weights on $\theta_{j,t-1}$ when calculating the best linear predictions are precisely the same weights as those computed for the empirical Bayes estimates.[35] See Appendix I for the proof that these two matrix expressions are equal.

While $(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}$ and $\big((\Omega + \Sigma_j)^{-1}\Omega\big)'$ are mathematically equivalent, there are

[34]The expectation here and in Equation (27) is a bit nuanced, as it is essentially is combining two conceptually different operations by both taking the expectation over the uncertain error terms as well as integrating over the population of teacher effects which are (in theory) fixed for each individual.

[35]This does not rule out the possibility that we can compute better non-linear predictors, even without the full set of normality assumptions. See Gilraine et al. (2020), for example.

reasons that using $\left((\Omega + \Sigma_j)^{-1}\Omega\right)'$ to compute the estimates is preferable. Most notably, writing $\Omega_j^* = (\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}$ requires that Ω is invertible. This assumption is violated if the set of measures can be summarized by a lower-dimension vector of true teacher effectiveness, such as their impact on students' cognitive and non-cognitive skills. Writing Ω_j^* as $\left((\Omega + \Sigma_j)^{-1}\Omega\right)'$, in contrast, no longer requires that Ω is invertible, and instead only that $\Omega + \Sigma_j$ is invertible. Note that even if Ω is theoretically invertible, it is possible that the estimates of $\hat{\Omega}$ will be not be invertible due to measurement error. Thus, estimation of Ω_j^* may be impossible when defining $\Omega_j^* = (\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}$ even if Ω is theoretically full rank. In contrast, this is not problematic when using the formulation that $\Omega_j^* = \left((\Omega + \Sigma_j)^{-1}\Omega\right)'$.

D Teachers with Missing Outcomes

Models for teacher value-added typically assume we observe noisy measures of teacher effectiveness for all of the outcomes we aim to predict. This is, however, often does not align with reality. For instance, we may want to estimate effectiveness for teachers in both tested and non-tested subjects/grades, but teachers in non-tested subjects/grades are missing test score measures. Similarly, future outcomes, such as grades and attendance, are not available in all years or for all grades. We now walk through how to estimate teacher effects when different teachers have different sets of observed measures.

D.1 Derivation of Empirical Bayes Posterior Distribution

First, we partition the full set of measures $\theta_{j,t-1}$ into unobserved measures, denoted $\theta_{1,j,t-1}$, and observed measures, denoted $\theta_{2,j,t-1}$. Similarly, we define $\Theta_{1,j}$ to be the true effects on the unobserved measures and $\Theta_{2,j}$ to be the true effects on the observed measures. We order the measures such that:

$$\theta_{j,t-1} = \begin{bmatrix} \theta_{1,j,t-1} \\ \theta_{2,j,t-1} \end{bmatrix} \qquad \text{and} \qquad \Theta_j = \begin{bmatrix} \Theta_{1,j} \\ \Theta_{2,j} \end{bmatrix}$$

but this ordering is without loss of generality and only for notational convenience.[36] We can similarly partition the covariance matrix of the true outcomes as: $\Omega = \begin{bmatrix} \Omega_{1,1} & \Omega_{1,2} \\ \Omega_{2,1} & \Omega_{2,2} \end{bmatrix}$ and the covariance matrix of the error terms as $\Sigma_j = \begin{bmatrix} \Sigma_{1,1,j} & \Sigma_{1,2,j} \\ \Sigma_{2,1,j} & \Sigma_{2,2,j} \end{bmatrix}$.

[36] In our implementation, we initially permute the measures to be in this form, use these equations to estimate the posterior mean and covariance, and then permute again to return the measures to their original ordering.

Since $\Theta_j \sim N\big(0, \Omega\big)$, we then get that:

$$\Theta_{1,j}|\Theta_{2,j} \sim N\Bigg(\Omega_{1,2}\Omega_{2,2}^{-1}\Theta_{2,j}, \Omega_{1,1} - \Omega_{1,2}\Omega_{2,2}^{-1}\Omega_{2,1}\Bigg) \tag{28}$$

We can therefore write that:

$$\Theta_j = \begin{bmatrix}\Omega_{1,2}\Omega_{2,2}^{-1} \\ I\end{bmatrix}\Theta_{2,j} + \eta \qquad \text{with} \qquad \eta \sim N\Bigg(0, \begin{bmatrix}\Omega_{1,1} - \Omega_{1,2}\Omega_{2,2}^{-1}\Omega_{2,1} & 0 \\ 0 & 0\end{bmatrix}\Bigg) \tag{29}$$

where I is the identity matrix with the number of rows equal to the number of measures the researcher observes. Similarly, we can the same derivation used to construct the empirical Bayes' estimates without missing data in Section II to get that:

$$\Theta_{2,j}|\theta_{2,j,t-1} \sim N\Bigg(\Omega_{2,2}\big(\Omega_{2,2} + \Sigma_{j,2,2}\big)^{-1}\theta_{2,j,t-1}, \big(\Omega_{2,2}^{-1} + \Sigma_{j,2,2}^{-1}\big)^{-1}\Bigg) \tag{30}$$

Again, we can use this expression to write $\Theta_{2,j}$ as a linear function of $\theta_{2,j,t-1}$ plus a normally distributed error term to get that:

$$\Theta_{2,j} = \Omega_{2,2}\big(\Omega_{2,2} + \Sigma_{j,2,2}\big)^{-1}\theta_{2,j,t-1} + \zeta \qquad \text{with} \qquad \zeta \sim N\Bigg(0, \big(\Omega_{2,2}^{-1} + \Sigma_{j,2,2}^{-1}\big)^{-1}\Bigg) \tag{31}$$

We can then plug in Equation (31) into (29) to get that:

$$\Theta_j = \begin{bmatrix}\Omega_{1,2}\Omega_{2,2}^{-1} \\ I\end{bmatrix}\Omega_{2,2}\big(\Omega_{2,2} + \Sigma_{j,2,2}\big)^{-1}\theta_{2,j,t-1} + \begin{bmatrix}\Omega_{1,2}\Omega_{2,2}^{-1} \\ I\end{bmatrix}\zeta + \eta \tag{32}$$

We then note that η is also independent from $\theta_{2,j,t-1}$, i.e., after conditioning the true effect of the teacher on the set of observed measures, the true effect of the teacher on the unobserved measures is independent from the estimated effects of the teacher on the observed measures. Thus, we can re-write Equation (32) as:

$$\Theta_j|\theta_{2,j,t-1} = N\Bigg(\begin{bmatrix}\Omega_{1,2} \\ \Omega_{2,2}\end{bmatrix}\big(\Omega_{2,2} + \Sigma_{j,2,2}\big)^{-1}\theta_{2,j,t-1}, \begin{bmatrix}\Omega_{1,2}\Omega_{2,2}^{-1} \\ I\end{bmatrix}Var(\zeta)\begin{bmatrix}\Omega_{1,2}\Omega_{2,2}^{-1} \\ I\end{bmatrix}' + Var(\eta)\Bigg) \tag{33}$$

and Equations (31) and (29) make clear that $Var(\zeta) = \big(\Omega_{2,2}^{-1} + \Sigma_{j,2,2}^{-1}\big)^{-1}$ and $Var(\eta) = \begin{bmatrix}\Omega_{1,1} - \Omega_{1,2}\Omega_{2,2}^{-1}\Omega_{2,1} & 0 \\ 0 & 0\end{bmatrix}$.

We conclude by emphasizing that there is an important assumption implicit in this approach, which is that after conditioning on $\theta_{2,j,t-1}$, the fact that we are missing $\theta_{1,j,t-1}$

tells us nothing about the underlying value of $\Theta_{1,j}$. This assumption would be violated if, for example, teachers are placed according to their comparative advantage. In that case, the fact that a teacher is slotted to teach a subject/grade without test scores is informative about her relative ability of improving students' test scores versus improving students' other outcomes. This assumption also underlies the implicit assumption that we know Ω or at least can estimate Ω precisely. Generally, we rely on individuals for whom we observe estimated effects on multiple outcomes to estimate the relationship between true teacher effects on each outcome (Ω). Thus, the key assumption here is that the relationship between the teacher effects on different outcomes is the same for teachers for whom all measures are observed as it is for the teachers for whom we only observe a subset of outcomes.

D.2 Comparison with Imputation Approach

The most natural alternative approach is to impute the missing values and then construct the empirical Bayes estimates according to Section II. Here we contrast the empirical Bayes approach outlined in Section D with imputation approaches.

To do so, we focus on two potential ways to impute the missing values. The easiest approach is to impute the missing values as $\mathbb{E}[\theta^k_{i,t}]$, if $\theta^k_{i,t}$ is the value that is missing. Note that the measures are normalized so that $\mathbb{E}[\theta^k_{i,t}] = 0$ for all k. Of course, this approach is problematic as it is does not distinguish between $\theta^k_{i,t}$ being missing and teacher i's impact on measure k as being average. Thus, the resulting empirical Bayes estimates are overly shrunken toward the mean. Note that since the empirical Bayes estimates of all measures will depend on $\theta^k_{i,t}$, the empirical Bayes estimates of all measures will be shrunken too much.

This comparison is a bit of a straw man, as we compare the method to the most simple imputation approach. A more complex imputation approach would be to impute the missing values as $\mathbb{E}[\theta^k_{i,t}|\theta^{-k}_{i,t}]$, where $\theta^{-k}_{i,t}$ is the set of measures which are not missing, before calculating the empirical Bayes estimates according to Section II. Note that $\mathbb{E}[\theta^k_{i,t}|\theta^{-k}_{i,t}]$ are themselves the empirical Bayes estimates, so among other things this approach is more complex to implement than the approach mentioned in Section D. It also means that, in some sense, the approach shrinks the estimates twice: first when constructing $\mathbb{E}[\theta^k_{i,t}|\theta^{-k}_{i,t}]$ and second when computing the empirical Bayes estimates post-imputation. This means that, while less obvious than the previous case, the resulting empirical Bayes estimates will be shrunken too much in this case as well.

There is, however, another force pushing this approach to shrink the empirical Bayes estimates too little. By imputing the missing values to $\mathbb{E}[\theta^k_{i,t}|\theta^{-k}_{i,t}]$, this approach assumes that we observe more information about individual i than we actually do. This alone would lead the empirical Bayes estimates are shrunken too little. Empirically, it appears the "double shrinkage" dominates and the resulting empirical Bayes estimates are indeed

shrunken to much.

To see this, we conduct a simulation where we randomly drop 25% of each observation and estimate the empirical Bayes estimates under three approaches: the missing value approach defined in Section D; the imputation approach where missing values are imputed to the overall mean; and the imputation approach where the missing values are imputed as $\mathbb{E}[\theta^k_{i,t}|\theta^{-k}_{i,t}]$. We then calculate each individuals' effectiveness, using the PCA Regression weights defined in Section II. Figure A.5 shows the three resulting distributions, focusing on individuals who are missing at least two observation. As can be seen, imputing the missing values at the overall mean shrinks the distribution much more than just treating the observations as missing. Similarly, imputing the missing values at $\mathbb{E}[\theta^k_{i,t}|\theta^{-k}_{i,t}]$ also shrinks the distribution more than just treating the observations as missing, although this is less pronounced than when imputing missing values to the overall mean.

Of course, the fact that the imputation approaches shrink the measures more than just treating is as missing does not alone mean that they are "overly shrunk" rather than the other distribution being under shrunken. In addition to the conceptual reasons to prefer the missing value approach over the imputation approaches, we can also provide some empirical evidence that it does better. While we do not observe the true effects, given our simulation we can compare the three empirical Bayes estimates to the empirical Bayes estimates generated when none of the observations are missing. When doing so, we find that the empirical Bayes estimates generated from the missing value approach are closer (as measured via mean-square error) to the ones when no variables are missing than the empirical Bayes estimates generated from either of the two imputation approaches.

E Conducting Principal Components Analysis with Multidimensional Empirical Bayes Estimates

Here, we walk through how to reduce the vector Θ_j of teacher j's K measures of effectiveness into a smaller vector of H measures, while minimizing loss of information about teacher j's effectiveness. Restricting our attention to linear transformations, we can express this transformation as a $K \times H$ matrix w, where the H measures of teacher effectiveness are $w'\Theta_j$.[37]

To formally define "information loss" we can reverse this transformation by taking the smaller vector of H measures, i.e., $w'\Theta_j$, and attempting to reconstruct the initial K measures. If we focus only on linear transformations, we can write this as $\tilde{w}(w'\Theta_j)$ for a $K \times H$

[37] As is clear from the proof, restricting ourselves to a linear transformation from Θ_j is not actually a restriction. Stated differently, the best rank-H approximation of Θ consists of a linear transformation that transforms the $N \times K$ data matrix to an $N \times H$ data matrix and then the "reversal," defined below, of this transformation to reconstruct a rank-H $N \times K$ data matrix.

matrix $\tilde{w}$. Since the initial linear transformation w maps a K dimensional space to an H dimensional space, it is impossible to perfectly reverse the transformation. One natural approach is to use the transpose of the initial matrix, i.e., $\tilde{w} = w$. If the initial transformation, for example, took an unweighted average of the K measures, then the transpose would map the average back to the K measures by setting each measure as $\frac{1}{K}$ times the average. Absent any additional information, this approach seems reasonable and there is indeed a mathematical justification for why that is the best approach.[38]

We can then define information loss as the difference between the true teacher effects on all K dimensions and the reconstructed teacher effects on the K dimensions, or $\sum_{\forall k}(\Theta_j^k - (ww'\Theta_j)^k)^2$.[39] Then, we can define the optimal weighting matrix ω^* as that which, given all j teachers, minimizes the information loss:

$$\omega^* = \arg\min_w \sum_{\forall j} \sum_{\forall k} (\Theta_j^k - (ww'\Theta_j)^k)^2 \tag{34}$$

While that seems like a challenging optimization problem, the first H components of a principal component analysis give the rows of ω^*. The intuition behind why this is true stems from the fact that the first component is the vector of weights that maximizes the variance of the resulting vector of data, which is essentially the same as minimizing the amount of remaining variance.[40] Since the remaining variance is the object we try to minimize in Equation (34), the first principal component is the optimal way to reduce the dimension of the data into a single dimension.[41]

The challenge here, and in many applied settings, is that we do not observe the "true" measures that we wish to use PCA to summarize. Rather, we have noisy estimates of the true measures and need to determine how to account for this noise in our principal components analysis. In particular, here we aim to summarize Θ, but we only observe the

[38]Formally, the transpose is connected to the inverse as follows: if the initial transformation is orthogonal and does not actually reduce the dimension, i.e., $H = K$, then the inverse of the initial transformation is the transpose, i.e., $ww' = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. If $H < K$, then w is the Moore-Penrose inverse of w'. Thus, w is the matrix such that $(w'w)w'\Theta_j = w'\Theta_j$ for every Θ_j.

[39]We include $ww'\Theta_j$ in parenthesis to emphasize that we apply ww' to the full vector Θ_j before taking the k^{th} measure.

[40]Formally defining the "amount of remaining variance" is a bit challenging, since the data has dimension K and the loadings that result from the first component have a single dimension. Without quite stating so explicitly, however, that is essentially what we discussed in the second paragraph of this section.

[41]For those interested in the technical details, the proof is as follows. An equivalent way to write Equation (34) is $\omega^* = \arg\min_w ||\Theta - \Theta ww'||_F$, where $||.||_F$ is the Frobenius norm. It is well-known that the best rank-H approximation to Θ, when using the Frobenius norm, is to conduct a singular value decomposition (SVD) on Θ and then use the H largest singular values and their corresponding singular vectors to construct a rank-H matrix. Let the SVD on Θ be written as $U\Sigma V'$, as is convention. Since V are the principal components, it then follows that $\Theta ww' = U\Sigma V' V_H V_H'$ when w consists of the first H principal components, denoted as V_H. Since the components are orthogonal, $V'V_H V_H' = V_H'$ and so $\Theta V_H V_H' = U\Sigma V_H' = U_H \Sigma_H V_H'$. Thus, $\Theta V_H V_H'$ the best rank-H approximation to Θ and so V_H is clearly the best choice of w.

value-added estimates, i.e., the $\theta_{j,t-1}$'s, and the empirical Bayes estimates, i.e., $\hat{\Theta}$.

We overcome this challenge by using the fact that the principal components correspond to the eigenvectors of the covariance matrix, Ω. More specifically, Ω can be factorized into $W\Lambda W^{-1}$, where W is the matrix of right eigenvectors and Λ is a diagonal matrix of eigenvalues. The columns of W are then the principal components, ordered in importance by the value of the corresponding eigenvalue, with the amount of variation explained by a component being equal to the value of its corresponding eigenvalue divided by the sum of the eigenvalues. In short, by using just the covariance matrix, we can estimate the H principal components with the largest eigenvalues to obtain the matrix of weights ω^* that solves the optimization problem defined in Equation (34). Applying the estimate ω^* to the matrix Θ gives the score matrix $\Theta\omega^*$, which is the best way to summarize Θ while using only H measures. Of course, while we can estimate ω^* without observing Θ, we cannot compute $\Theta\omega^*$ without observing Θ. Thus, we also need to compute the empirical Bayes' estimates of $\Theta\omega^*$ (in the same way we computed the empirical Bayes estimates of Θ). We denote this as $\hat{\Theta}\omega^*$, where $\hat{\Theta}$ are the empirical Bayes estimates of Θ.[42]

F Using Empirical Bayes' Estimates as Covariates

Researchers often use the empirical Bayes estimates as covariates in a subsequent regression. In cases where the empirical Bayes estimate consist of a single dimension and are the only covariate in this regression, it is well known that one can interpret the coefficient as if the true measure was used in the regression (Jacob and Lefgren (2008)). We show here that the same is true when the empirical Bayes estimates are multidimensional and when other covariates are included in the regression; however, there are some subtitles that we discuss as well. We start by presenting the formal statement and proof, then discuss the result, and then provide some empirical results that illustrate how subtle issues in the implementation can impact that resulting coefficients.

F.1 Formal Statement and Proof

To formalize this, suppose that we want to use the empirical Bayes estimates as regressors, i.e., we want to estimate a regression of some outcome $\tilde{\Theta}_j$ on Θ_j. We will let γ be the OLS coefficient resulting from that regression, i.e.,:

$$\gamma = \lim_{N\to\infty} (\Theta'\Theta)^{-1}\Theta'\tilde{\Theta} \tag{35}$$

[42]The fact that the empirical Bayes' estimates of $\omega^*\Theta$ are $\omega^*\hat{\Theta}$ follows from the fact that if a $m \times 1$ vector x is distributed normally $N(\mu, \Sigma)$, then $w'x \sim N(w'\mu, w'\Sigma w)$ for any $m \times 1$ vector of weights w.

where the j^{th} row of Θ is Θ_j'. Since we do not observe Θ_j directly, however, we instead need to estimate

$$\hat{\gamma} = \lim_{N\to\infty} (\hat{\Theta}'\hat{\Theta})^{-1}\hat{\Theta}'\tilde{\Theta} \tag{36}$$

where the j^{th} row of $\hat{\Theta}$ is $\hat{\Theta}_j'$ and $\hat{\Theta}_j = \Omega_j^*\theta_{j,t-1}$. Give these definitions, we can present the formal theorem:

Theorem 1. *Using the definitions above,* $\hat{\gamma} = \gamma$.

Proof. We start by using the law of large numbers, together with the fact that $\hat{\Theta}_j = \Omega_j^*\theta_{j,t-1}$, to get that $\frac{1}{N}\hat{\Theta}'\hat{\Theta} \to \mathbb{E}[(\Omega_j^*\theta_{j,t-1})(\Omega_j^*\theta_{j,t-1})']$. From the assumptions inherent to the the model we discuss in Section II.A, it follows that $\mathbb{E}[(\Omega_j^*\theta_{j,t-1})(\Omega_j^*\theta_{j,t-1})'] = \mathbb{E}[\Omega_j^*(\Omega+\Sigma_j)\Omega_j^{*'}]$. Thus, $\frac{1}{N}\hat{\Theta}'\hat{\Theta} \to \mathbb{E}[\Omega_j^*(\Omega+\Sigma_j)\Omega_j^{*'}]$.

Next, also using the law of large numbers we get that $\frac{1}{N}\hat{\Theta}'\tilde{\Theta} \to \mathbb{E}[(\Omega_j^*\theta_{j,t-1})\tilde{\Theta}_j]$. From the fact that γ is the OLS coefficient resulting from a regression of $\tilde{\theta}_j$ on Θ_j, we can write $\tilde{\Theta}_j = \Theta_j'\gamma + e_j$, where $\mathbb{E}[\Theta_j e_j] = 0$. Thus, $\mathbb{E}[(\Omega_j^*\theta_{j,t-1})\tilde{\Theta}_j] = \mathbb{E}[(\Omega_j^*\theta_{j,t-1})\Theta_j'\gamma + e_j] = \mathbb{E}[\Omega_j^*\theta_{j,t-1}\Theta_j']\gamma + \mathbb{E}[\Omega_j^*\theta_{j,t-1}e_j]$.

We will assume that $\mathbb{E}[\theta_{j,t-1}e_j] = 0$ since $\mathbb{E}[\Theta_j e_j] = 0$, which is essentially assuming that the estimation error for Θ_j is uncorrelated with the outcome of interest $\tilde{\Theta}_j$. This would generally be true if, for example, the long-run outcome of interest is measured using a different cohort of students than is used to estimate the short-term impact.

Under this assumption, we get that $\frac{1}{N}\hat{\Theta}'\tilde{\Theta} \to \mathbb{E}[\Omega_j^*\Omega]\gamma$. This follows from the fact that $\mathbb{E}[\theta_{j,t-1}\Theta_j'] = \Omega$, which reflects the fact that the estimation error is uncorrelated with the true impact of the teacher.

Combining the above two results, we get that:

$$\hat{\gamma} = \mathbb{E}[\Omega_j^*(\Omega+\Sigma_j)\Omega_j^{*'}]^{-1}\mathbb{E}[\Omega_j^*\Omega]\gamma \tag{37}$$

which itself implies that $\hat{\gamma} = \gamma$ if (and only if) $\mathbb{E}[\Omega_j^*(\Omega+\Sigma_j)\Omega_j^{*'}] = \mathbb{E}[\Omega_j^*\Omega]$. Using the formulation that $\Omega_j^* = (\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}$ it is far from obvious that this is the case. That it is true, however, is easy to see when using the alternative description, that $\Omega_j^* = \Omega(\Omega+\Sigma_j)^{-1}$. From this, we get that:

$$\begin{aligned}
\mathbb{E}[\Omega_j^*(\Omega+\Sigma_j)\Omega_j^{*'}] &= \mathbb{E}\Omega(\Omega+\Sigma_j)^{-1}(\Omega+\Sigma_j)(\Omega+\Sigma_j)^{-1}\Omega] \\
&= \mathbb{E}[\Omega(\Omega+\Sigma_j)^{-1}\Omega] \\
&= \mathbb{E}[\Omega_j^*\Omega]
\end{aligned}$$

□

F.2 Discussion of Results

An important implication of this proof is that using the "correct" Ω_j^*, i.e. $\Omega_j^* = \Omega(\Omega+\Sigma_j)^{-1}$, is not only important for efficiency reasons (e.g. the difference between weighted least squares and ordinary least squares), but is a requirement for the consistency of the resulting coefficients. Stated differently, using a different Ω_j^* leads to inconsistent coefficient estimates, i.e. $\hat{\gamma} \neq \gamma$.[43] While this is clear from the proof, this has a number of important implications. First, it is worth noting that conducting the empirical Bayes' shrinkage separately for each measure corresponds to a different Ω_j^* and therefore would lead to inconsistent coefficients in any resulting regression.[44]

More subtly, suppose after estimating the empirical Bayes' estimates on a number of short-term measures, one first ran a series of simple linear regressions to look at how each measure individual was related to the long-term outcome of interest before then running a regression that included all of the empirical Bayes' estimates in a single regression. Confusingly, while the coefficients from the final regression could be interpreted as if the true measures were used as covariates in this case, the coefficients from the simple linear regressions could not be interpreted this way. If one wants to conduct this analysis, the above result suggests that one should estimate conduct the empirical Bayes' shrinkage differently for each regression that is run where the set of measures used to construct Ω_j^*, and hence the empirical Bayes' estimates, is restricted to those used in the regression.

Similarly, suppose that either to improve identification or precision, one hopes to include additional covariates in the regression of the long-term outcome on the empirical Bayes' measures. Again, the above results suggest that unless the additional covariates are uncorrelated with both the true effects and the measurement error, the resulting coefficient estimates will be inconsistent.[45]

All of these points are more apparent when the use of empirical Bayes' estimates as covariates is viewed as the second stage of a two-stage least squared approach to dealing with measurement error, in which $\theta_{j,t}$ serve as instruments for Θ_j.[46] One subtle difference

[43] To see this, take the simply example where every teacher has the same number of students, in which case Ω_j^* is identical for all j and so we can ignore the expectations. Thus, $\Omega_j^*(\Omega + \Sigma_j)\Omega_j^{*'} = \Omega_j^*\Omega$ can be solved directly to get that $\Omega_j^{*'} = (\Omega+\Sigma_j)^{-1}\Omega$. When teachers have different number of students, it becomes more complicated and the Ω_j^* required for consistency is no longer unique: most notably both $(\Omega+\Sigma_j)^{-1}\Omega$ and $\mathbb{E}[(\Omega+\Sigma_j)^{-1}\Omega]$ would work. This subtlety does not impact the points discussed below, however.

[44] More specifically, conducting the empirical Bayes' shrinkage separately for each measure corresponds to an Ω_j^* that is identical to $\Omega(\Omega+\Sigma_j)^{-1}$ on the diagonals and zero everywhere else. This is therefore only identical to $\Omega(\Omega+\Sigma_j)^{-1}$ if both the true effects and the measurement error are uncorrelated across measures.

[45] To see this, we can think of simply extending $\hat{\Theta}$ to include these covariates. This changes Ω and Σ_j, but does not change the fact that Ω_j^* needs to equal $\Omega(\Omega+\Sigma_j)^{-1}$. Unless Ω and Σ_j are both block diagonal matrices, with the blocks corresponding (at least) to the $\theta_{j,t-1}$'s and the additional covariates, one cannot do the empirical Bayes' only on the $\theta_{j,t-1}$'s and still obtain the correct result.

[46] One important implication of this is that it suggests that it would be possible to leverage results from

is that in our context the coefficients from the "first stage" differ depending on the number of students the teacher has, which is why we present the formal proof in the appendix. This can be seen most clearly when one views Ω_j^* as the coefficients from a regression of Θ_j on $\theta_{j,t-1}$, which we discuss in the text. This means that $\hat{\Theta}_j$ are the predicted values from this regression, which can then be plugged in to the second-stage regression.

F.3 Empirical Example

In the discussion above, we highlighted that we needed to use the multivariate empirical Bayes approach – as opposed to generating an empirical Bayes estimates separately for each measure – when using multiple measures in a regression. Here, we present some empirical results highlighting this fact.

To do so, we assume that there are two measures in which we are interested, Θ_1 and Θ_2, which are distributed $N(0, \Omega)$ with $\Omega = \left(\begin{smallmatrix} 1 & \rho_\Omega \\ \rho_\Omega & 1 \end{smallmatrix}\right)$. The measures themselves are estimated with error, which itself is distributed $N(0, \Sigma_j)$ with $\Sigma_j = \left(\begin{smallmatrix} 1 & \rho_\epsilon \\ \rho_\epsilon & 1 \end{smallmatrix}\right)$.

Finally, we assume that outcome of interest is related to the two measures of interest as follows:

$$\tilde{\Theta}_j = 0.5\Theta_{1,j} + 1.0\Theta_{2,j} + \nu_j \tag{38}$$

where ν_j is an error term that is independent from both Θ_1 and Θ_2 as well as the error with which they are estimated.

We then consider two regressions: one in which the first measure is the only covariate and one in which both measures are included as covariates. For each regression, we then consider three ways in which the convariates are constructed. In the first, they are estimated via the multidimensional empirical Bayes approach described here; in the second, the single-dimensional empirical Bayes approach is run separately for each measure. Finally, we also include a case where the true measures are observed without error, which serves as the benchmark.

The results are presented in Table A1 below, which only show the coefficient on the first measure for both regressions to simplify the presentation of results.

As seen under the columns under the header "Single Linear Regression," when the first measure is the only one included in the subsequent regression constructing the empirical Bayes using only that measure produces coefficients identical to the case where the true measures are observed, while constructing the empirical Bayes using both measures as described in this paper produces different coefficients depending on the correlation of the two measures' error terms. When the correlation in the error term is less than the correlation

the research on errors-in-variables if one was interested in estimating non-linear relationships between the true teacher effects and an outcome of interest (e.g., Amemiya (1985); Hausman et al. (1991); Hong and Tamer (2003); Lewbel (1998); Hu and Schennach (2008)).

		Single Linear Regression			Multiple Linear Regression		
ρ_Ω	ρ_ϵ	Multi EB	Single EB	Truth	Multi EB	Single EB	Truth
0.5	-0.9	1.80	1.0	1.0	0.5	1.13	0.5
0.5	-0.7	1.71	1.0	1.0	0.5	1.07	0.5
0.5	-0.5	1.63	1.0	1.0	0.5	1.0	0.5
0.5	-0.3	1.52	1.0	1.0	0.5	0.92	0.5
0.5	-0.1	1.41	1.0	1.0	0.5	0.83	0.5
0.5	0.0	1.36	1.0	1.0	0.5	0.78	0.5
0.5	0.1	1.30	1.0	1.0	0.5	0.73	0.5
0.5	0.3	1.16	1.0	1.0	0.5	0.63	0.5
0.5	0.5	1.0	1.0	1.0	0.5	0.5	0.5
0.5	0.7	0.82	1.0	1.0	0.5	0.36	0.5
0.5	0.9	0.62	1.0	1.0	0.5	0.19	0.5

Table A1: This table shows how the coefficients in a single linear regression and multiple linear regression depend on whether the empirical Bayes estimates are constructed via a multidimensional empirical Bayes method (Multi EB) or separately (Single EB) and how those coefficients relate to the coefficients obtained for a regrssion where the true measures were observed (i.e., Truth).

between the true measures, the coefficient converges to a parameter larger than the true coefficient, i.e., the coefficient obtained if one was to observe the true measure.

In contrast, as seen under the columns under the header "Multiple Linear Regression," when both measures are included in the regression it is important to construct the empirical Bayes estimates using the multidimensional approach. While doing so always gives the same coefficient on the first measure as it would if the true measure was observed, if the empirical Bayes measures are instead constructed separately for each measure the resulting coefficient depends on how the covariance in the error term relates to the covariance between the true measures. Again, if the covariance of the error terms is less than the covariance between the true measures, the resulting coefficient is overestimated and vice versa. While note shown, the coefficient on the second measure in the multivariate regression has a similar pattern, in which the coefficient is overestimated when the coveriance of the error terms is less than the covariance of the true measures and underestimated when the covariance of the error term is more than the covariance of the true measures.

G Implied Weights on the Raw Effect Estimates

Note that there were two sets of weights that we discussed in Section V. The first is the set of weights implied by the multidimensional empirical Bayes that turn the combined raw

estimates into the best estimates of the teachers' true effects, denoted by Ω_j^*.[47] The second is the set of weights that determine how a principal can reduce the multiple dimensions of teacher effectiveness into a small number of summary measures, e.g., the first eigenvalue of the short-term effectiveness measures or their relative relationship with long-term effectiveness measures. Here, we combine the two results to illustrate the weights the different raw estimates receive when computing the final measures.

The key is to leverage the fact that $\mathbb{E}[\Theta_j|\theta_{j,t-1}] = \Omega_j^*\theta_{j,t-1}$, where as before Ω_j^* is the matrix implied by the multidimensional empirical Bayes approach and is defined above in Section II. As a reminder of notation, Θ_j corresponds to the true effectiveness of teacher j and $\theta_{j,t-1}$ is the raw estimate of teacher effectiveness, i.e., the average residuals as opposed to the empirical Bayes' value-added estimates. It follows that $\mathbb{E}[\omega'\Theta_j|\theta_{j,t-1}] = \omega'\Omega_j^*\theta_{j,t-1}$ for any set of weights ω that one wants to put on the true measures of effectiveness. Thus, $\omega'\Omega_j^*$ are the weights on the raw measures, which we present below.

As in the previous section we focus on three potential choices for ω:

1. First Eigenvalue: Use the vector of weights from first principal component.

2. PCA Regression: Use the coefficients from a regression of high school graduation rates on empirical Bayes' estimates of the first four principal components.

3. Regression: Use the coefficients from a regression of high school graduation rates on empirical Bayes' estimates of the K outcomes.

Table A.4 uses the PCA and regression results to construct these three types of weights for elementary and middle school teachers. Note that the specific weights depend on Ω_j^*, which varies across teachers and depends on how many students they taught.[48] For our example, we focus on a hypothetical teacher who teachers the average number of students. Columns one to three contain the unstandardized weights, while the weights in columns four to six are standardized according to the variance in teacher effects on the relevant outcome. Thus, columns one to three give the weights that should actually be used on the raw outcomes (i.e. $\omega'\Omega_j^*$), while columns four to six illustrates how important each of the raw outcomes are in determining the summative measure.

For elementary school, (panel (A) of Table A.4), teacher effects on future outcomes receive a lot more weight than teacher effects on current test scores (and attendance). Weights on attendance are typically small and always negative. The weights on current test scores vary across the weighting approach employed and in the PCA regression approach, the weights on math test scores are negative.

[47] Explicitly, the estimates are "best" under a mean-squared loss function and the normality assumptions.

[48] In the case where multiple years of data are incorporated into the empirical Bayes' measures, it will also depend on how many years teachers are in the data.

For middle school, (panel (B) of Table A.4), teacher effects on future grades in subjects other than those taught receive the most weight. Test scores also receive substantial weight. The relative weights of the four remaining dimensions vary across methods.

Which set of weights they will want to use depends on the goals of evaluation and what underlying measure of effectiveness the decision maker is trying to summarize. The weights from the regressions in columns (2) and (3) are likely most appropriate when the decision maker cares about placing the most weight on the short-term measures most related to longer-term outcomes.[49] The weights from the first eigenvalue, in contrast, are more appropriate when the decisionmaker simply aims to best summarize effects on the short-term outcomes.

H Incorporating Multiple Years of Data into the Estimates

H.1 Without Drift

In the model without drift in teacher effectiveness, incorporating multiple years of data into the estimates is straightforward. This is because the assumption of no drift in effectiveness implies the teacher effect estimates in year $t-2$, i.e. $\theta_{j,t-2}$, are just as predictive of teacher effectiveness in year t, i.e. $\Theta_{j,t}$, as are the teacher effect estimates from year $t-1$, i.e. $\theta_{j,t-1}$. We therefore do not need to distinguish between $\theta_{j,t-1}$, $\theta_{j,t-2}$, etc. and instead can just condition on the average of the teacher effect estimates.

Formally, suppose that teacher j has been in the data for M years prior to year t. We can then define:

$$\overline{\theta}_{j,-t} = \sum_{m=1}^{M} \theta_{j,t-m} \tag{39}$$

Under the assumption of no drift, we can use the same derivation as before to get an almost identical expression:

$$\mathbb{E}\left[\Theta_{j,t}|\overline{\theta}_{j,-t}\right] = \Omega^{*}\overline{\theta}_{j,-t} \tag{40}$$

where as before $\Omega^{*} = \left(\Sigma_j^{-1} + \Omega^{-1}\right)^{-1}\Sigma_j^{-1}$, Ω is the covariance matrix of the true teacher effects and Σ_j is the covariance matrix of the error terms implicit in $\overline{\theta}_{j,-t}$. The only additional challenge here is to estimate Σ_j now that the empirical Bayes' estimate is conditioning on an average measure over years (and students within each year) as well as over students in

[49] The differences between columns (2) and (3) is less a question of what the decision maker cares about and more a practical question of whether reducing the dimensions of the data before the regression helps improve the predictions.

a single year. From the assumptions discussed in Section II.A, it follows that:

$$\Sigma_j = \frac{1}{M}\Sigma_\nu + \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N_{j,t-m}}\Sigma_\epsilon \tag{41}$$

where $N_{j,t-m}$ is the number of students teacher j taught in year $t-m$.[50]

As we discuss in Appendix B, the assumption of no drift in teacher effectiveness is not particularly consequential when including only a single year in the empirical Bayes estimates. However, whether one allows for drift in teacher effectiveness does impact the interpretation and estimation of the empirical Bayes estimates when multiple years are included in the estimates. Intuitively, this is because drift in teacher effectiveness means the estimated teacher effects from year $t-1$ are more predictive of the teacher's effect in year t than the estimated teacher effects from year $t-M$. Thus, when constructing the posterior distribution, one should give more weight to the estimates from year $t-1$ than on the ones from year $t-M$. Appendix H explains this in more depth and shows how one can compute the empirical Bayes' estimates of multidimensional teacher quality in a model with drift in teacher effectiveness.

H.2 With Drift

We next use the model presented in Appendix B to construct the empirical Bayes' estimates in a model which allows for drift in teacher effectiveness.

To do so, we will initially focus on the case where we only aim to condition on two years, $\theta_{j,t-1}$ and $\theta_{j,t-2}$, rather than the more general case of conditioning on M years. It is easy to see how this can be extended to the more general case.

To start, we note that:

$$\begin{pmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{pmatrix}\Bigg|\begin{pmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix} \sim N\left(\begin{bmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{bmatrix},\begin{bmatrix}\Sigma_{j,t-1} & 0\\ 0 & \Sigma_{j,t-2}\end{bmatrix}\right) \tag{42}$$

where $\Sigma_{j,t-1} = \Sigma_\nu + \frac{1}{N_{j,t-1}}\Sigma_\epsilon$ and $N_{j,t-1}$ is the number of students teacher j teaches in year $t-1$. Most notably, once you condition on $\Theta_{j,t-1}$ and $\Theta_{j,t-2}$, $\theta_{j,t-1}$ and $\theta_{j,t-2}$ are independent.

[50] We subtly jumped to conditioning on $\overline{\theta}_{j,-t}$ rather than on $\theta_{j,t-1}, \theta_{j,t-2}, ..., \theta_{j,t-M}$. In a model without drift, this is mostly inconsequential, although it is not actually quite optimal. Instead, one should condition on a weighted average of the previous estimates, with the weights being proportional to the variance of the estimates. In practice, we expect (and encourage) researchers and practitioners to allow for drift in teacher effectiveness when using multiple years of data to construct teacher value-added estimates. We outline how to do so in Appendix H. If one wants to use the optimal weights without allowing for drift, one can rely on the results presented here and assume that the covariances between the years are all identical.

Next, from our assumptions on drift, we get that

$$\begin{pmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix}\right) \tag{43}$$

Thus, from Bayes' Law we get that:

$$\mathbb{E}\left[\begin{pmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix} \middle| \begin{pmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{pmatrix} \right] = \begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix} \begin{bmatrix}\Omega+\Sigma_{j,t-1} & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega+\Sigma_{j,t-2}\end{bmatrix}^{-1} \begin{bmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{bmatrix} \tag{44}$$

Finally, from our assumptions on drift we get that:

$$\begin{pmatrix}\Theta_{j,t}\\ \Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix} \sim N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\Omega & \mathbb{C}_1 & \mathbb{C}_2\\ \mathbb{C}_1 & \Omega & \mathbb{C}_1\\ \mathbb{C}_2 & \mathbb{C}_1 & \Omega\end{bmatrix}\right) \tag{45}$$

and so

$$\Theta_{j,t}\middle| \begin{pmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix} \sim N\left(\begin{bmatrix}\mathbb{C}_1 & \mathbb{C}_2\end{bmatrix} \begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix}^{-1} \begin{bmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{bmatrix}, \mathbf{\Sigma}\right) \tag{46}$$

for a covariance matrix $\mathbf{\Sigma} = \Omega - \begin{bmatrix}\mathbb{C}_1 & \mathbb{C}_2\end{bmatrix} \begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix}^{-1} \begin{bmatrix}\mathbb{C}_1\\ \mathbb{C}_2\end{bmatrix}$. Thus, we get that

$$\begin{aligned}\mathbb{E}\Big[\Theta_{j,t}| \begin{pmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{pmatrix}\Big] &= \begin{bmatrix}\mathbb{C}_1 & \mathbb{C}_2\end{bmatrix} \begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix}^{-1} \begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix} \begin{bmatrix}\Omega+\Sigma_{j,t-1} & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega+\Sigma_{j,t-2}\end{bmatrix}^{-1} \begin{bmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{bmatrix}\\ &= \begin{bmatrix}\mathbb{C}_1 & \mathbb{C}_2\end{bmatrix} \begin{bmatrix}\Omega+\Sigma_{j,t-1} & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega+\Sigma_{j,t-2}\end{bmatrix}^{-1} \begin{bmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{bmatrix}\end{aligned}$$

I Additional Proofs

Theorem 2. *Let* $\Omega = \left(\begin{smallmatrix}\sigma^2_{\Omega,1} & \rho_\Omega\\ \rho_\Omega & \sigma^2_{\Omega,2}\end{smallmatrix}\right)$ *and* $\Sigma_j = \left(\begin{smallmatrix}\sigma^2_{\Sigma,1} & \rho_\Sigma\\ \rho_\Sigma & \sigma^2_{\Sigma,2}\end{smallmatrix}\right)$. *If denote* $\Omega^*_j = \left(\begin{smallmatrix}\omega_{1,1} & \omega_{1,2}\\ \omega_{2,1} & \omega_{2,2}\end{smallmatrix}\right)$, *we get that:*

$$\omega_{1,1} = \frac{1}{det(\Omega+\Sigma_j)}\left[\sigma^2_{\Omega,1}\sigma^2_{\Omega,2} + \sigma^2_{\Omega,1}\sigma^2_{\Sigma,2} - \rho^2_\Omega - \rho_\Omega\rho_\Sigma\right] \tag{47}$$

$$\omega_{1,2} = \frac{1}{det(\Omega+\Sigma_j)}\left[\sigma^2_{\Sigma,1}\rho_\Omega - \sigma^2_{\Omega,1}\rho_\Sigma\right] \tag{48}$$

Proof. This is most clearly seen using the fact that Ω^*_j can also be written as $\left((\Omega+\Sigma_j)^{-1}\Omega\right)'$,

which we prove below. We then get that:

$$\Omega_j^* = \left((\Omega + \Sigma_j)^{-1}\Omega\right)' \tag{49}$$

$$= \left[\frac{1}{det(\Omega + \Sigma_j)} \begin{pmatrix} \sigma_{\Omega,2}^2 + \sigma_{\Sigma,2}^2 & -(\rho_\Omega + \rho_\Sigma) \\ -(\rho_\Omega + \rho_\Sigma) & \sigma_{\Omega,1}^2 + \sigma_{\Sigma,1}^2 \end{pmatrix} \begin{pmatrix} \sigma_{\Omega,1}^2 & \rho_\Omega \\ \rho_\Omega & \sigma_{\Omega,2}^2 \end{pmatrix}\right]' \tag{50}$$

$$\tag{51}$$

where $det(\Omega + \Sigma_j)$ is the determinant of $\Omega + \Sigma_j$. Multiplying the matrices and accounting for the transpose, we get that:

$$\omega_{1,1} = \frac{1}{det(\Omega + \Sigma_j)}\left[(\sigma_{\Omega,2}^2 + \sigma_{\Sigma,2}^2)\sigma_{\Omega,1}^2 - \rho_\Omega(\rho_\Omega + \rho_\Sigma)\right] \tag{52}$$

$$\omega_{1,2} = \frac{1}{det(\Omega + \Sigma_j)}\left[(\sigma_{\Omega,2}^2 + \sigma_{\Sigma,2}^2)\rho_\Omega - \sigma_{\Omega,2}^2(\rho_\Omega + \rho_\Sigma)\right] \tag{53}$$

$$= \frac{1}{det(\Omega + \Sigma_j)}\left[\sigma_{\Sigma,2}^2\rho_\Omega - \sigma_{\Omega,2}^2\rho_\Sigma\right] \tag{54}$$

$$\omega_{2,1} = \frac{1}{det(\Omega + \Sigma_j)}\left[(\sigma_{\Omega,1}^2 + \sigma_{\Sigma,1}^2)\rho_\Omega - \sigma_{\Sigma,1}^2(\rho_\Omega + \rho_\Sigma)\right] \tag{55}$$

$$= \frac{1}{det(\Omega + \Sigma_j)}\left[\sigma_{\Sigma,1}^2\rho_\Omega - \sigma_{\Omega,1}^2\rho_\Sigma\right] \tag{56}$$

$$\omega_{2,2} = \frac{1}{det(\Omega + \Sigma_j)}\left[(\sigma_{\Omega,1}^2 + \sigma_{\Sigma,1}^2)\sigma_{\Omega,2}^2 - \rho_\Omega(\rho_\Omega + \rho_\Sigma)\right] \tag{57}$$

□

Theorem 3. *For any symmetric, invertible matrices Σ_j and Ω such that $\Sigma_j + \Omega$ is also invertible, we have:*

$$(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1} = \left((\Omega + \Sigma_j)^{-1}\Omega\right)' \tag{58}$$

Proof. We first note that if two matrices A and B are invertible, then $A = B$ if and only if $A^{-1} = B^{-1}$. So we will show that $\left[(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}\right]^{-1} = \left[\left((\Omega + \Sigma_j)^{-1}\Omega\right)'\right]^{-1}$. Using the properties of inverses, we get that:

$$\left[(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}\right]^{-1} = \Sigma_j(\Sigma_j^{-1} + \Omega^{-1}) \tag{59}$$

$$= \Sigma_j\Sigma_j^{-1} + \Sigma_j\Omega^{-1} \tag{60}$$

$$= \mathbf{I} + \Sigma_j\Omega^{-1} \tag{61}$$

where $\mathbf{I}$ is the identity matrix.

Similarly, using the properties of inverses, transposes, and the the fact that Ω and Σ_j

are symmetric, we get that:

$$\left[\left((\Omega+\Sigma_j)^{-1}\Omega\right)'\right]^{-1} = \left[\Omega(\Omega+\Sigma_j)^{-1}\right]^{-1} \tag{62}$$

$$= (\Omega+\Sigma_j)\Omega^{-1} \tag{63}$$

$$= \mathbf{I}+\Sigma_j\Omega^{-1} \tag{64}$$

Two final notes. First, the condition that Σ_j and Ω are both invertible, as is $\Sigma_j+\Omega$, is satisfied when Σ_j and Ω are positive definite matrices. Thus, the conditions for the proof will hold in our context as long as Ω is invertible. Second, this proof provides yet another way to express the weights, where $\Omega_j^* = \left(\mathbf{I}+\Sigma_j\Omega^{-1}\right)^{-1}$. This also makes clear that the weights depend on the relative size of the error terms, Σ_j, and the true effects, Ω. □

Theorem 4. *Under the estimate approach specified in Section II.A, the model specified in Section II.C, and the assumption that the number of teachers increases to infinity and each teacher has multiple students, we have that:* $\hat{\beta} \to \beta$, $\hat{\Omega} \to \Omega$, $\hat{\Sigma}_\epsilon \to \Sigma_\epsilon$, $\hat{\Sigma}_\nu \to \Sigma_\nu$, *and* $\hat{\Sigma}_\phi \to \Sigma_\phi$.

Proof. We start by showing that under the assumptions, $\hat{\beta} \to \beta$ as the number of students goes to infinity. To do so, we note that from the Frisch-Waugh-Lovell theorem including teacher fixed effects is equivalent to demeaning the outcome and covariates at the teacher-level and then running a regression at the student-level without the teacher fixed-effects. Denoting $\overline{X}_{j,t}$ as the average outcome on measure X over the students who teacher j teaches in year t, our statistical model of student outcomes (e.g. Equation (19)) implies that:

$$y_{i,t} - \overline{y}_{j,t} = \beta \cdot (X_{i,t} - \overline{X}_{j,t}) - (\phi j'(i), t - \overline{\phi}_{j,t}) - (\epsilon_{i,t} - \overline{\epsilon}_{j,t}) \tag{65}$$

Under our assumption that the error terms are distributed normally and independent from the other variables, the coefficient from the regression of $y_{i,t-1} - \overline{y}_{j,t-1}$ on $X_{i,t-1} - \overline{X}_{j,t-1}$ will converge to β as the number of students go to infinity.

From this, we get that $\hat{\beta} = \beta + o_p(1)$ and so we can ignore differences between $\hat{\beta}$ and β when considering consistency. It therefore follows that $\hat{e}_{i,t} = \phi_{i,t} + \epsilon_{i,t}$ and therefore $\hat{\Sigma}_\epsilon \to \Sigma_\epsilon$.

Similarly, the fact that $\hat{\beta} = \beta + o_p(1)$ implies that asymptotically $\theta_{j,j',t} = \Theta_j + \nu_{j,t} + \phi_{j'(i),t} + \frac{1}{N_j}\sum \epsilon_{i,t}$. From this and our assumptions on the independence of the error terms and that fact that $\hat{\Sigma}_\epsilon \to \Sigma_\epsilon$, it follows that $\hat{\Omega} \to \Omega$ when $\hat{\Omega}$ is defined by Equation (**??**).

We can prove that $\hat{\Sigma}_\nu \to \Sigma_\nu$ and $\hat{\Sigma}_\phi \to \Sigma_\phi$ in a similar fashion, i.e., by using assumptions on the independence of the error terms and the fact that our previous estimated parameters converge to their true values. □