ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-451

Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness

Christine Mulhern
RAND Corporation

Isaac M. Opper
RAND Corporation

There is an emerging consensus that teachers impact multiple student outcomes, but it remains unclear how to summarize these multiple dimensions of teacher effectiveness into simple metrics that can be used for research or personnel decisions. Here, we discuss the implications of estimating teacher effects in a multidimensional empirical Bayes framework and illustrate how to appropriately use these noisy estimates to assess the dimensionality and predictive power of the true teacher effects. Empirically, our principal components analysis indicates that the multiple dimensions can be efficiently summarized by a small number of measures; for example, one dimension explains over half the variation in the teacher effects on all the dimensions we observe. Summary measures based on the first principal component lead to similar rankings of teachers as summary measures weighting short-term effects by their prediction of long-term outcomes. We conclude by discussing the practical implications of using summary measures of effectiveness and, specifically, how to ensure that the policy implementation is fair when different sets of measures are observed for different teachers.

VERSION: August 2021

Suggested citation: Mulhern, Christin, and Isaac M. Opper. (2021). Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness. (EdWorkingPaper: 21-451). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/h9qh-0078

Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness*

Christine Mulhern Isaac M. Opper

August 18, 2021

Abstract

There is an emerging consensus that teachers impact multiple student outcomes, but it remains unclear how to summarize these multiple dimensions of teacher effectiveness into simple metrics that can be used for research or personnel decisions. Here, we discuss the implications of estimating teacher effects in a multidimensional empirical Bayes framework and illustrate how to appropriately use these noisy estimates to assess the dimensionality and predictive power of the true teacher effects. Empirically, our principal components analysis indicates that the multiple dimensions can be efficiently summarized by a small number of measures; for example, one dimension explains over half the variation in the teacher effects on all the dimensions we observe. Summary measures based on the first principal component lead to similar rankings of teachers as summary measures weighting short-term effects by their prediction of long-term outcomes. We conclude by discussing the practical implications of using summary measures of effectiveness and, specifically, how to ensure that the policy implementation is fair when different sets of measures are observed for different teachers.

*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, and Amherst College 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' impact, 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 body of work now documents that teacher effects extend beyond traditional measures of test score effects, with teachers influencing outcomes such as attendance and student behavior (Gershenson (2016); Jackson (2018); Kraft (2019); Liu and Loeb (2019); Petek and Pope (2018)). Furthermore, teachers who are effective at increasing test scores are not necessarily effective at improving socio-emotional skills, so traditional test score value-added does not necessarily identify the "best" teachers.

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 can be used for personnel decisions. This is especially challenging given that not all measures are observed for all teachers; for example, not all teachers teach in grades with an end-of-year test. In this paper, we discuss the challenges and implications of estimating teacher value-added in a multidimensional framework, develop approaches for constructing summary measures of teacher effects, and present results on the dimensionality of teacher effects and summary measures. We also propose a new way to use value-added measures in practice which ensures that teachers are not unfairly advantaged or disadvantaged based on which set of outcomes or measures are observed.

We consider two broad approaches for summarizing teacher effects when their effects are multidimensional. These approaches seek to balance decisionmakers' goals of identifying teachers' true effectiveness with practical evaluation limitations. First, we consider how to optimally reduce the dimensions of short-term effectiveness on which teachers are evaluated

[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.

while minimizing information loss. For this, we use principal components analysis of teacher effects on short-term outcomes to estimate dimensions of teacher effects and create summary measures. Second, we consider a case where a decisionmaker wants to evaluate teachers based on their long-term effects. For this, we examine how to weight short-term measures of effectiveness to optimally predict long-term effectiveness.

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, the 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. We therefore start with a discussion of the practical implications of using 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, there are several important implications for estimation and inference that deserve discussion and which help in the interpretation of our empirical results.

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 will incorporate information about the estimates of all the other dimensions. This means that the best estimate of a teacher's impact on test scores, for example, 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. Whether the test score empirical Bayes estimate is shrunk towards or away from the attendance estimate depends on both the correlation of the true effects and the correlation of their 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 on attendance if the error terms are also positively correlated.

We then discuss how to use estimates from the multidimensional empirical Bayes frame-

work to derive estimates of 1) the principal components, and 2) the relationship between the short-term effects and long-term effects, which converge to the same parameters as those generated from the true measures of teacher effectiveness. In this discussion, we explain why doing principal components analysis on the empirical Bayes estimates is different (and less preferred) than computing empirical Bayes estimates of the principal components. We also show that multidimensional empirical Bayes estimates can be used as regressors to uncover the true relationships between the measures and outcomes of interest, a fact that is well-known in the single dimension case where the empirical Bayes is a shrunken version of the noisy estimate.

We apply these techniques to estimate and summarize the effects of thousands of New York City teachers. In doing so, we find three main empirical results. First, we show that more than half of the variation in teacher effects on the outcomes we observe can be captured with one dimension. Four dimensions capture nearly all the variation in teacher effects on the six to eight outcomes we examine. While the first dimension is essentially an average of all observed outcomes, the second dimension separates teacher effects on grades from effects on test scores. The remaining two dimensions are more difficult to interpret, although for elementary school teachers one of the dimensions appears to separate effects on math achievement from those on English Language Arts (ELA) achievement. We also show that the short-term measures can explain 27% to 48% of the variation in teacher effects on their students' probability of graduating high school on time.

Second, we examine three approaches for combining short-term measures of effectiveness to create summary measures and show that teacher ratings are very similar across these approaches. For this, we consider three ways to weight the short-term measures. These weights are based on 1) the eigenvalue of the first principal component, 2) coefficients from a regression of high school graduation rates on the four main principal components, and 3) coefficients from a regression of graduation rates on estimates of teacher effects on each individual outcome. Each dimension of effectiveness (or student outcome) receives similar weight across these three approaches. Weights, and resulting summary measures, are, how-

ever, sensitive to properly accounting for noise in observed teacher measures; ignoring the noise with which teacher teacher effects are estimated or the covariance of the measures of teacher effects leads to noticeably different estimates.

Third, we show that teacher rankings across the three summary measures are very highly correlated. They are also highly correlated with teacher rankings from traditional non-test score measures of value-added. While test score value-added measures are positively correlated with these other measures of teacher effects, there is noticeable information loss when only relying on test scores. In addition, there is little overlap in the teachers who are at the bottom 5% in terms of the summary measures and test score value-added.[2] Thus, which measures are used for evaluation can have important implications for individual teachers.

Finally, we discuss extensions that allow for computing value-added when different measures are observed for different teachers. In particular, we show how one can compute measures of teacher effectiveness on a set of dimensions even if the teacher is missing data for those outcomes using information about the teacher's estimated effects on other outcomes and the estimated covariance of teacher effects across all dimensions. In practice this is often necessary, as the likelihood that one can estimate teacher effects on all dimensions diminishes as the number of dimensions increases. While this approach allows researchers to generate value-added estimates on the full set of outcomes, using these outcomes for policy without acknowledging their uncertainty will generally result in policies that unfairly advantage or disadvantage teachers depending on how many measures are observed. We therefore conclude by illustrating that explicitly accounting for the uncertainty in the value-added estimates enables one to guarantee that the resulting policy is fair.

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

[2]This is consistent with existing research, e.g., Jackson (2018) and Petek and Pope (2018).

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 the teachers' effect on non-test score outcomes are more predictive of the teachers' effect on students' long-term outcomes than the teachers' effect on test scores (e.g., 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. In writing the paper, the authors gained a much better understanding of how the empirical Bayes estimates are constructed in a multidimensional setting and we hope that in reading the paper, other researchers do as well.

The paper proceeds as follows: Section 2 presents the conceptual framework and data; Section 3 describes the analytic approach; Section 4 presents the empirical Bayes estimates and the results on how well the effects can be summarized with a lower dimensional vector; Section 5 discusses the practical implications the results have for teacher evaluation; Section 6 focuses on the implementation challenges that occur when not all measures are observed for all teachers; Section 7 concludes.

II Conceptual Framework and Data

II.A Conceptual Framework

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, post-secondary attainment, and earnings (Bacher-Hicks et al. (2019, 2020); Chamberlain (2013); Chetty et al. (2014a,b); Gershenson (2016); Gershenson et al. (2018); Jackson (2018); Kraft (2019); Ladd and Sorensen (2017)). We denote the effect that teacher j would have on student i's k^{th} outcome in year t if she taught him as $\Theta^k_{j,t}$ and the full vector of effects as $\Theta_{j,t}$. Thus, if we randomly switched student i from teacher j to teacher j', we would expect his k^{th} outcome to change by $\Theta^k_{j',t} - \Theta^k_{j,t}$.

While quite general, we note two assumptions implicit in this formulation. First, by not indexing this effect by i, we assume teacher j has the same effect on all her potential students (Delgado (2020); Aucejo et al. (2020)). Second, the specification treats $\Theta_{j,t}$ as fixed under the status quo. In practice, a teacher's effectiveness on each dimension is not an innate characteristic, and will depend on the teacher's pre-teaching training, the context in which they teach (e.g. school climate, leadership, and on-the-job training), and both the explicit and implicit incentives they face (Taylor and Tyler (2012); Aucejo et al. (Forthcoming); Dee and Wyckoff (2015); Macartney (2016); Papay et al. (2020); Rockoff et al. (2012)).

We refer to $\Theta^k_{j,t}$ as the teacher's effect on outcome k. If the true dimensionality of teacher effectiveness is less than K, it is possible to summarize $\Theta_{j,t}$ by some lower dimensional vector of teacher effectiveness. This would be the case if teachers' effects can, for example, be grouped into effects on students' cognitive skills and effects on students' non-cognitive skills. How well $\Theta_{j,t}$ can be summarized by a lower dimensional vector is one of the empirical questions we answer in this paper.

Next, suppose there is a principal who needs to make some personnel decisions in year $t-1$, such as whether to require teacher j to get professional development. Naturally, she

wants to incorporate information about each teachers' effect on the various outcomes into her decision. Doing so is challenging, however, because she cannot observe $\Theta_{j,t}$ directly. Instead, at best she observes a noisy measure of their effectiveness. We denote the k^{th} measure of *measured* teacher effectiveness of teacher j in year $t-1$ as $\theta^k_{i,t-1}$ and denote the full vector of measured teacher effectiveness as $\theta_{j,t-1}$. We discuss in the next section how $\theta_{j,t-1}$ is measured and how best to use a multidimensional empirical Bayes' framework to improve on the raw estimates of teacher effectiveness.

In addition, she does not observe noisy measures of effectiveness for some of the outcomes about which she cares. For example, while she might care about teachers' effects on high school graduation, she will not observe these effects in a timely manner for any of her teachers. Thus, in this paper, we also explore 1) how to combine the measures of short-term effectiveness if the decision-maker cares only about long-term effectiveness and 2) how much information is lost by not directly observing effects on the long-term outcome.

Finally, there is a third challenge in that not all measures of effectiveness are observed for all teachers. For example, many teachers do not teach in tested grades and thus do not have traditional test score value-added measures. In addition, if the principal wants to incorporate multiple years of data into the decision, or use measures based on student outcomes in future years, the amount of information available for each teacher will depend on how long she has been teaching in the district. Since principals need to make personnel decisions for all teachers, it is necessary to find a fair approach to evaluation, which does not arbitrarily punish or reward teachers based on how many measures are observed. We discuss this challenge and offer a potential solution in Section VI.

II.B Data and Setting

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 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 absences, 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 in 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 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.

III Multidimensional Empirical Bayes Estimation

III.A Model Details and Intuition

This section describes a multidimensional empirical Bayes' estimation strategy 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, which enables error terms to be correlated across measures within a year. This leads to different variance estimates than if value-added measures within years are assumed to be independent. After presenting the framework and deriving the results, we briefly discuss computational considerations and the resulting weights.

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 denote student i's k^{th} outcome in year t as $y_{i,t}^k$ and the full vector of student i's outcomes in year t as $y_{i,t}$. We let $X_{i,t}$ be a vector of p student covariates and take the estimate $\hat{\beta}$ as fixed. Like most of the value-added literature, we assume that student outcomes can be expressed as a linear function of: their teacher's effect on their outcome; a vector of their covariates; a classroom-level shock shared by all students denoted as $\tilde{\nu}_{j,t}$; and an individual level shock denoted as $\epsilon_{i,t}$.

Thus, our statistical model of student outcomes is:

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

Next, we model how $\Theta_{j,t-1}$, teachers' true effects in year $t-1$, relate to $\Theta_{j,t}$, teachers' true effects in year t. For now, we do not incorporate drift in teacher effectiveness as it complicates the model presentation. Instead, we assume that a teacher's effect on their students' outcomes is a combination of the teacher's persistent effectiveness and a year-specific shock to their effectiveness, $\Theta_{j,t-1} = \Theta_j + \eta_{j,t-1}$, for some persistent effect Θ_j and a year specific shock $\eta_{j,t}$. As we show in Appendix C, extending the model to account for drift does not change in the interpretation of our results. Determining whether or not to account for drift does matter, however, if the principal aims to incorporate multiple years of data into their year t predictions; we discuss this in Appendix F.

One benefit of assuming away drift is that we can define a new error term $\nu_{j,t-1} = \eta_{j,t-1} + \tilde{\nu}_{j,t-1}$. One can think of $\nu_{j,t-1}$ as an error term that combines the classroom shock that is not caused by the teacher (embedded in the $\tilde{\nu}_{j,t-1}$ term) with the classroom-level shock that is caused by the teacher, but not related to a teacher's persistent effectiveness (embedded in the $\eta_{j,t-1}$ term).[3] 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 the principal's decision discussed in Section II.A.[4]

The statistical model of student outcomes thus becomes:

$$y_{i,t-1} = \beta X_{i,t-1} + \Theta_j + \nu_{j,t-1} + \epsilon_{i,t-1} \tag{2}$$

Our key assumption is that the two error terms are independently distributed across teachers and years, normally distributed, and have mean zero, so $\nu_{j,t-1} \sim N(0, \Sigma_\nu)$ and $\epsilon_{i,t-1} \sim N(0, \Sigma_\epsilon)$. The assumption that the error terms are independently distributed across years means that while value-added measures are noisy, they are also unbiased.

[3]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.

[4]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 C, separating them also is important when one allows for teacher drift.

Thus, 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 (Chetty et al. (2014a); Bacher-Hicks et al. (2019); Petek and Pope (2018)), especially when the measures are test scores. As we discuss below, assuming the error terms are normal is less important, though relaxing this assumption means the estimates no longer correspond to the mean of the Bayes posterior distribution. Finally, while we assume the errors are independent across years, we do not assume that they are independent across measures within a year. This distinguishes 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)).

We then define a teacher's value-added in year $t-1$ as:

$$\theta_{j,t-1} = \frac{1}{N_j} \sum_{\forall i \in C(j,t-1)} y_{i,t-1} - \hat{\beta} X_{i,t-1} \tag{3}$$

where $C(j, t-1)$ is the set of students teacher j teaches in year $t-1$ and $||C(j, t-1)|| = N_j$.

From this statistical model, we get that:

$$\theta_{j,t-1} | \Theta_j \sim N\Big(\Theta_j, \Sigma_\nu + \frac{1}{N_j}\Sigma_\epsilon\Big) \tag{4}$$

under the assumption that $\hat{\beta} \approx \beta$. We further assume that teachers' persistent effectiveness is normally distributed with $\Theta_j \sim N(0, \Omega)$. Bayes' Law then implies that:

$$\Theta_j | \theta_{j,t-1} \sim N\Big(\Omega_j^* \theta_{j,t-1}, \Sigma_j^*\Big) \tag{5}$$

where

$$\Sigma_j = \Sigma_\nu + \frac{1}{N_j}\Sigma_\epsilon$$
$$\Omega_j^* = (\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}$$
$$\Sigma_j^* = (\Sigma_j^{-1} + \Omega^{-1})^{-1}$$

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 $\hat{\Theta}_j$.

This empirical Bayes' framework relies on the assumption that both the true teacher effects and the error terms are normally distributed; however, the normality assumptions are less important than one might expect. This is because the empirical Bayes estimates are also equivalent to the best linear predictors of the true teacher effects given the previous years' estimated teacher effects, a fact that is true even if the error terms and/or true effects are not normally distributed. 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:[5]

$$\Psi_j^{k*} = \arg\min_{\Psi^k} \mathbb{E}\Big[\big(\Theta_j^k - \Psi^k\theta_{j,t-1}\big)'\big(\Theta_j^k - \Psi^k\theta_{j,t-1}\big)\Big] \tag{6}$$

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:

$$\Psi_j^{k*} = \mathbb{E}\Big[\Big((\theta_{j,t-1}'\theta_{j,t-1})^{-1}\theta_{j,t-1}'\Theta_j^k\Big)'\Big] \tag{7}$$

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 $(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1} = \big((\Omega + \Sigma_j)^{-1}\Omega\big)'$.

[5]The expectation here and in Equation (7) 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.

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.[6] See Appendix B for the proof that these two matrix expressions are equal.

While $(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}$ and $\left((\Omega + \Sigma_j)^{-1}\Omega\right)'$ are mathematically equivalent, there are 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 $\Omega_j^* = \left((\Omega + \Sigma_j)^{-1}\Omega\right)'$.

Understanding the Empirical Bayes Estimates: In the multidimensional setting, the matrix Ω_j^* contains the weights that are used to translate the various measures of measured teacher quality, $\theta_{j,t-1}$, 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.[7] For example, the empirical Bayes' estimate of a teacher's ability to improve students' attendance would likely include information on the estimated ability of the teacher to improve student test scores as well as the estimated ability of the teacher to improve student attendance.

[6] 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.

[7] Note this statement is not quite true, as it is possible that the weighted covariance of Σ_j is equal to the weighted covariance of Ω in which case 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 clear in the example below.

To build intuition on how the empirical Bayes estimates incorporate information from the various measures, we now walk through an example with two outcomes. To do so, we let $\Omega = \begin{pmatrix} \sigma^2_{\Omega,1} & \rho_\Omega \\ \rho_\Omega & \sigma^2_{\Omega,2} \end{pmatrix}$ and $\Sigma_j = \begin{pmatrix} \sigma^2_{\Sigma,1} & \rho_\Sigma \\ \rho_\Sigma & \sigma^2_{\Sigma,2} \end{pmatrix}$. Here ρ_Ω and ρ_Σ correspond to the covariance between the two true measures of teacher effectiveness and two error terms, respectively, rather than the correlation between the measures. If we denote $\Omega^*_j = \begin{pmatrix} \omega_{1,1} & \omega_{1,2} \\ \omega_{2,1} & \omega_{2,2} \end{pmatrix}$, we get:[8]

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

$$\omega_{1,2} = \frac{1}{det(\Omega + \Sigma_j)} \left[\sigma^2_{\Sigma,1}\rho_\Omega - \sigma^2_{\Omega,1}\rho_\Sigma \right] \tag{9}$$

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.[9]

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

[8] See Appendix B for proof.

[9] While we use the term "correlated" here, Equation (18) 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.

with an average teacher.[10] The relative variance and covariances of the true effects versus the error terms inform us whether a) or b) is the more likely explanation, 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.

III.B Summarizing Teacher Effectiveness

The conceptual framework and estimation details lay out an approach for estimating teacher effectiveness on all K measures. That is, the empirical Bayes estimates, $\hat{\Theta}_j$, are the best predictors of the true teacher effects, Θ_j, given the raw teacher residuals, $\theta_{j,t-1}$. While these estimates may be the best predictors of true teacher effectiveness, that still leaves us with K measures of teacher quality when principals and researchers may want or need to summarize teacher effectiveness using fewer dimensions. We next discuss three possible approaches to efficiently summarize Θ_j and the complications in their implementation due to the fact that we do not observe Θ_j directly.

Principal Component Analysis: One natural approach to reduce the number of measures is principal components analysis (PCA), which is commonly used to reduce the dimensionality of a dataset in a way that minimizes information loss. We aim to reduce the vector Θ_j of teacher j's K measures of effectiveness into a smaller vector of H measures, while losing the minimum amount 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 $w'\Theta_j$.[11]

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 $K \times H$

[10] This analysis assumes that the 's 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.

[11] 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.[12]

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 $\sum_{\forall k}(\Theta_j^k - (ww'\Theta_j)^k)^2$.[13] Then, we can define the optimal weighting matrix ω^* as that which, given all j teachers, minimizes the information loss:

$$\omega^* = \arg\min_w \sum_{\forall j} \sum_{\forall k} (\Theta_j^k - (ww'\Theta_j)^k)^2 \tag{10}$$

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.[14] Since the remaining variance is the object we try to minimize in Equation (10), the first principal component is the optimal way to reduce the dimension of the data into a single dimension.[15]

[12]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.

[13]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.

[14]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.

[15]For those interested in the technical details, the proof is as follows. An equivalent way to write Equation (10) 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

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 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. Thus, as long as we can consistently estimate Ω, we can estimate the principal components of Θ. Importantly, as we discuss below, we can estimate Ω even though do not observe Θ directly.

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 (10). 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.[16]

Finally, it is worth noting that the empirical Bayes' estimates of the score matrix, i.e., $\hat{\Theta}\omega^*$, do not generally give the same results as conducting a PCA on the empirical Bayes'

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, $V'V_HV_H' = V_H'$ 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.

[16]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.

estimates directly, i.e., conducting a PCA on $\hat{\Theta}$.[17] 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."

Relationship of Short-Term Measures to Long-Term Outcomes: Another natural approach is to view the short-term effectiveness measures as proxies, or statistical surrogates, for the effect the teacher has on their students' long-term outcomes, such as high school graduation or lifetime earnings (Petek and Pope (2018); Chetty et al. (2014b)). When the short-term measures are viewed as surrogates, we care about the predicted effect of the teacher on the long-term outcome given the vector of the teacher's short-term effects, rather than the full vector of teacher effects. This also reduces the dimensions of teacher 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 and $\tilde{\theta}_j$ to be the estimate of the teachers' impact on this outcome. For simplicity, we will assume there is a single long-term outcome of interest.[18] We then define:

$$\omega^* = \arg\min_{\omega} \frac{1}{J} \sum_{\forall j} (\tilde{\Theta}_j - \omega' \Theta_j)^2 \tag{11}$$

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 $\tilde{\Theta}_j$.

One natural approach is to estimate ω^* by replacing Θ with $\hat{\Theta}$, i.e., the matrix of empirical Bayes estimates, and $\tilde{\Theta}_j$ with $\tilde{\theta}_j$, i.e., the true long-term effect with the estimated long-term effect, to get $\hat{\omega}^* = (\hat{\Theta}'\hat{\Theta})^{-1}\hat{\Theta}'\tilde{\theta}$. It is well known that $\omega^* = \hat{\omega}^*$ in settings where Θ_j is one-dimensional, e.g., Jacob and Lefgren (2008). Thus, in the one-dimensional setting, one can use the empirical Bayes estimates as covariates and interpret the coefficient as if the true teacher effect was the covariate. It is not, however, obvious from Jacob and Lefgren's

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

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

(2008) proof in the single dimension that this extends to the multidimensional framework because the empirical Bayes estimates for each dimension consist of linear combinations of all the dimensions, rather than simply being the shrunken version of the estimates (which the Jacob and Lefgren (2008) proof assumes). In Appendix B we prove that the result extends to the multidimensional case and that $\omega^* = \hat{\omega}^*$ even in cases where where Θ_j is multidimensional.[19] Thus, as before one can use the multidimensional empirical Bayes estimates as covariates to uncover the relationship between the dependent variable and true teacher effects.

The key idea behind the proof is that using the empirical Bayes estimates as regressors is essentially the second-stage of a 2SLS regression, in which the $\theta_{j,t-1}$ serve as instruments for Θ_j. The idea is therefore essentially the same as using IV to adjust for measurement error, when presented with multiple noisy measures of the same underlying variable.[20] One subtle difference 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 above. This means that $\hat{\Theta}_j$ are the predicted values from this regression, which can then be plugged in to the second-stage regression.

One subtlety of the proof is that it relies on the assumption that the same set of measures are used to estimate the value-added as appear in the subsequent regression.[21] For example, suppose that we observe math and ELA test scores and use these two measures to estimate value-added using the multidimensional empirical Bayes' approach. Then, suppose we run three regressions: one that only includes the math value-added as a regressor; one that

[19]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 III.A.

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

[21]It is also worth acknowledging explicitly that the proof also assumes that no additional covariates are included in the regression.

only includes ELA value-added as a regressor; and one that includes both. The proof in Appendix B 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.*

In addition to estimating ω^*, a natural question is whether the set of teachers' short-term measures are sufficient to explain most of the variation in the teachers' long-term effect. Again, this would be straightforward if the true effect of the teachers were observed, on both the short-term and long-term outcomes. In that case, one would simply use the R^2 measure from a regression of the $\tilde{\Theta}_j$ on Θ_j. Since we have noisy measures of teacher effects, we instead use the year-to-year covariance of the estimated long-term effect to estimate the true variance of the long-term effect.[22] The fact that we do not directly observe the short-term measures also complicates our analysis, since the variance explained by the value-added estimates also differs from the (hypothetical) variance explained by the true-measures.[23]

[22]This is similar to what we did for the principal components analysis. As before, the assumption required for this to work is that teachers are not consistently assigned students who do better or worse on their long-term outcomes than would be expected based on their prior test scores, demographics, etc. If there is sorting on the long-term measures - a very real possibility - the true variance of the long-term effect would be biased upward. As is clear below, this would bias downward our estimates of how much of the long-term effect variation the short-term measures can explain.

[23]The fact that we do not directly observe the short-term measures also complicates our analysis because even if teachers' true short-term measures could explain all of the variation in teachers' long-term effects, the teachers' *estimated* short-term measures may not explain most of the variation. The percent of the

Denoting estimated variance as $\hat{\sigma}^2_{LT}$, the percent of the long-term variance explained by the true short-term measures, is $\frac{\omega^{*'}\Omega\omega^*}{\hat{\sigma}^2_{LT}}$, where $\omega^{*'}$ are the coefficient estimates. 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.

Combining the Approaches: In the previous discussion, we implicitly assumed that ω^* is well-defined. This in turn assumes Θ is full rank and not able to be summarized by a smaller dimension of teacher effects. Even if Θ is full rank, if some of K measures of teacher effectiveness are highly correlated, the estimates of $\hat{\omega}^*$ would be quite imprecise given that $\Omega'\Omega$ would then be nearly invertible.

A natural solution is to combine the two approaches to dimensionality reduction. In this, we first conduct principal components analysis to reduce the K dimensions into H components. We then include the empirical Bayes estimates of these H dimensions as covariates in a regression with the long-run outcome of interest as the dependent variable.

III.C Estimation Details

We estimate teacher effects and the matrices described above using data from 3rd to 8th grade students in New York City. Estimation involves the following five steps. Appendix B contains the proof that this approach provides consistent estimates of the relevant matrices.

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

long-term variance explained by the *estimated* short-term measures is quite similar to the result below: $\frac{\omega^{*'}(\Omega(\Omega+\Sigma_j)^{-1}\Omega)\omega^*}{\hat{\sigma}^2_{LT}}$. The difference reflects the fact that the variance of the empirical Bayes' estimates is $\Omega(\Omega+\Sigma_j)^{-1}\Omega$, while the variance of the true measures is simply Ω.

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.**

$$\hat{\Sigma}_\epsilon = \frac{1}{N}\sum_{\forall i} \hat{e}_{i,t}\hat{e}'_{i,t} \tag{12}$$

where $\hat{e}_{i,t} = y_{i,t} - \hat{\beta}X_{i,t} + \hat{\zeta}_j$.

3. **Calculate $\theta_{j,t}$ using the estimates of $\hat{\beta}$ from step 1.**

$$\theta_{j,t-1} = \frac{1}{N_j}\sum_{\forall i \in C(j,t-1)} y_{i,t-1} - \hat{\beta}X_{i,t-1} \tag{13}$$

where $C(j, t-1)$ is the set of teacher j's students in year $t-1$ and $N_j = ||C(j, t-1)||$.

4. **We estimate Ω using the fact that $\mathbb{E}[\theta_{j,t}\theta'_{j,t-1}] = \Omega$.** We can set:

$$\hat{\Omega} = \frac{1}{J}\sum \theta_{j,t}\theta'_{j,t-1} \tag{14}$$

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

5. **Back out Σ_ν using the fact that $\mathbb{E}[\theta_{j,t}\theta'_{j,t}] = \Omega + \Sigma_\nu + \frac{1}{N_j}\Sigma_\epsilon$.** We therefore estimate Σ_ν as:

$$\hat{\Sigma}_\nu = \frac{1}{J}\sum \theta_{j,t}\theta'_{j,t} - \hat{\Omega} - \frac{1}{N_j}\hat{\Sigma}_\epsilon \tag{15}$$

where J is the number of teacher-years and N_j is the number of students assigned to teacher j.

Given these estimates, we can also compute the full error covariance matrix, i.e., $\hat{\Sigma}_j = \hat{\Sigma}_\nu + \frac{1}{N_j}\hat{\Sigma}_\epsilon$, 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., $\hat{\Theta}_j = \hat{\Omega}^*_j\theta_{j,t-1}$.

IV Empirical Bayes Estimates of Teacher Effectiveness and Dimensionality Results

This section describes our teacher effect estimates and the degree to which these can be summarized by a lower dimension. We first focus on elementary school teachers and then on middle school teachers. In doing so, we focus on the dimensionality of teachers' true effects, rather than the estimated effects. While we do not observe these directly, we can use the covariance matrix of teacher effects to understand how they relate to one another. As discussed above, this enables us to describe the underlying dimensions of teacher effects and, theoretically, how a principal who cares about teacher effects on long-term outcomes would want to weight measures of short-term effectiveness if she directly observed the teachers' true short-term impacts. In Appendix E, we discuss how to use these weights with observed measures of teacher effects, which are the weights a principal or researcher would need to use in practice.

IV.A Elementary School

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.

Figure 1 summarizes the empirical Bayes estimates for each of these outcomes. Here, we have standardized the outcomes at the student level, rather than the value-added measures themselves. Thus, 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.[24]

From the empirical Bayes estimates alone, it is impossible to know how much varia-

[24]It is certainly possible that this is due to attendance being loosely tracked in the administrative data during our time span. That explanation would not, however, explain why there is seemingly variation in future attendance.

tion 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 eight outcomes. Although we cannot directly observe true effects, we can estimate the correlations as implied from our estimate of Ω. Effects on math and ELA tests are very highly correlated, with a coefficient of 0.72. Teacher effects on current tests are also moderately correlated with their effects on future tests, with coefficients between 0.33 and 0.55. Teacher effects on current test scores, however, are less correlated with effects on attendance and grades.

IV.A.1 Dimensions of Effectiveness

Next, we use principal components analysis to assess the dimensionality of teacher effects for elementary school. Panel (A) of Figure 3 shows the proportion of variance explained by each of the principal components (the values are reported in Table A.1). The first component

explains 49% of the variation, and the first four components collectively explain over 91% of the variance. These results indicate that our initial eight dimensions of effectiveness can be reduced to a smaller set of dimensions without losing much information. We focus on the first four components since together they explain over 90% of the variance and individually explain at least 5% of the variance.[25]

Table 4 and Figure 4 show the composition of the four main principal components in terms of the eight original outcomes. 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. The fourth component negatively weights ELA test scores effects and positively weights effects on math tests scores and future attendance. Attendance receives very little weight in all of these components.

IV.A.2 Long Term Predictions of Value-Added Dimensions

Next, we look at how these principal components, as well as our eight empirical Bayes estimates of effectiveness, relate to teachers' long-term effects on high school graduation.[26] In other words, we use the short-term measures as statistical surrogates for the long-term outcome of interest (Prentice (1989); Athey et al. (2019); Begg and Leung (2000)). To do so, 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

[25]The columns of Table A.1 highlight the importance of our methods in conducting principal components 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.

[26]In the Appendix, we also look at long-term effects on earning a Regents diploma and advanced Regents diploma.

year $t-1$). This ensures that the error term of the outcome is uncorrelated with the error of the empirical Bayes estimates.

These estimates are useful for thinking about how to weight the dimensions of teacher effects when constructing a summary measure if the evaluator primarily cares about the long-term effect of the teacher. 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.

Column (1) of Table 5 shows that the first component of teacher effectiveness is also the most predictive of high school graduation. Thus, if the goal is to identify teachers' whose effects are most related to high school graduation, the first component should receive the most weight. The second component also receives significant weight. These coefficients indicate that a one standard deviation improvement in teacher effectiveness in terms of component one leads to a 2.6 percentage point increase in graduation rates, conditional on effectiveness on the other components, and a one standard deviation improvement in teacher effectiveness in terms of component two leads to a 1.3 percentage point decrease in graduation rates. Note that since the second component mostly distinguishes between current test scores and future grades, the negative coefficient reflects the fact that future grades are deemed more important than current test scores and that the component (arbitrarily) gave positive weight to current test scores and negative weight to future grades. So "improvement" in component two means larger impact on future grades and smaller impact

on current test scores. The third component is not predictive of high school graduation and, while marginally statistically significant, the fourth component is only weakly related. The results are similar when predicting receipt of a Regents diploma (Table A.3). For advanced Regents diplomas, the third component is also a significant predictor.

Panel (B) of Table 5 shows the relationships between the individual empirical Bayes estimates and long-term outcomes, conditional on the other effectiveness measures. Here, future attendance and future math grades are most predictive of high school graduation. These coefficients, however, do not indicate the individual predictive power of effectiveness measures. Rather, they are the association between dimension K of effectiveness and high school graduation conditional on the other $k-1$ measures of effectiveness.[27] Thus, these coefficients can be thought of as weights one may want to place on individual measures of effectiveness if the goal is to identify teachers effective at improving high school graduation.

Finally, the bottom row of Table 5 reports that although the short-term measures can explain a non-trivial fraction of the long-term teacher effectiveness, nearly three-quarters of the total variation in long-term elementary school teacher effectiveness is unexplained by their effectiveness on current and future test scores, attendance, and grades.

IV.B Middle School

Next, we repeat the above analyses but for middle school teachers, who teach grades 6 and 7.[28] Most middle school teachers in our sample do not teach both math and English, so we focus on subject-specific outcomes. This gives us six potential outcomes over which we can construct teacher value-added: test scores in subject taught, future test scores, attendance, future attendance, future grades in subject, and future grades in other subjects. Figure 2 shows the distribution of the empirical Bayes estimates for middle school teachers.

[27] The empirical Bayes methods we discussed earlier are meant to be used in settings like this where all empirical Bayes estimates are collectively used in a regression. If we instead wanted to regress high school graduation on one measure of effectiveness, we should construct different empirical Bayes measures using the one-dimensional setup as in Kane & Staiger (2008).

[28] We omit 8th grade teachers, since we do not have future test scores, attendance, or grades for their students.

Once again, there is little variation in teacher effects on current attendance, but much larger variation in their effects on test scores and grades. Panel (B) of Table 3 shows the correlations of the empirical Bayes estimates. Teacher effects on current and future test scores are very highly correlated, with a coefficient of 0.89. Test score effects are positively correlated with effects on future grades but barely correlated with effects on attendance. Thus, teachers who improve student attendance are often not the same as those who improve test scores and grades.

IV.B.1 Dimensions of Teacher Effectiveness

Panel (B) of Figure 3 shows the principal components analysis for middle school teachers. Here, component one explains 68% of the variation, component two explains 16%, and the first four components collectively explain 99% of the variation in the six outcomes. Thus, we can use a lower-dimensional measure of effectiveness than the six dimensions we started with without losing much information. As before, we focus on the first four components as they each explain at least 5% of the variation.

Panel (B) of Table 4 describes how each of the six outcomes are weighted in each of the four main principal components. This can also be seen graphically in panel (B) of Figure 4. The first component is primarily based on teacher effects on future grades, though effects on test scores receive some positive weight. Component two separates teacher effects on test scores from effects on attendance and grades. Component three heavily weights effects on future attendance, while component four appears to differentiate between future grades on the same subject and future grades on different subjects. Effects on current attendance barely contribute to any of the four main components.

IV.B.2 Long-Term Predictions of Value-Added Dimensions

Next, Table 5 shows how related these short-term measures of effectiveness are to teachers' longer-term effects on high school graduation. This is useful for thinking about how much weight to put on the PCA components or individual measures of effectiveness if we care

about identifying middle school teachers who improve high school graduation.

Column 4 of Table 5 shows that components one and two are significantly and positively related to graduation. Component one is the most positively related to high school graduation, with a one standard deviation improvement in teacher effectiveness on this dimension, conditional on the other components, increasing graduation by 2.1 percentage points.[29] For the individual empirical Bayes estimates, teacher effects on future grades in other subjects (i.e., subjects outside that taught by teacher j) are most related to high school graduation. The coefficients on attendance and future attendance are also both statistically significant, though they are of opposite signs. This highlights the difficulty in interpreting the coefficients, as one might wonder what it means for a teacher to improve current attendance and then lower future attendance. Since these coefficients represent the predictive power of the individual dimension conditional on all the other measures of effectiveness, they are most useful for thinking about ways to weight components when creating summary measures, rather than for analyzing the predictive power of individual dimensions.

As shown in the bottom row of Table 5, the short-term measures can explain a larger fraction of variation in long-term effectiveness for middle school teachers than for elementary school teachers. Specifically, for middle school teachers the about one-half of the variation in long-term effectiveness can be explained by their effectiveness on current and future test scores, attendance, and grades.

V Implications for Teacher Evaluation

In practice, the goal of school principals and policymakers is often to identify the most (or least) effective teachers. The results from the previous section can be used to create weighted summary measures of teacher effects. We focus on summary measures based on the following three types of weights.

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

[29] The results are similar when looking at Regents diplomas and advanced Regents diplomas (Table A.3).

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 6 and Figure 5 summarize how the three summary measures are constructed based on these weights and observed outcomes. Columns one to three contain the unstandardized weights applied to each measure of teacher effectiveness, 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. Appendix E describes the construction of these weighted summary measures in more detail.

While the weights placed on individual measures clearly vary across these three approaches, it is not clear how much the set of most (or least) effective teachers identified differs across these approaches. The first three columns of Table 7 show that teachers ratings are highly correlated across the three weighting approaches, both in elementary and middle school. The correlations range from 0.90 for the eigenvalue weights and regression weights in elementary school, to 0.99 for the eigenvalue and regression weights in middle school. Figures 6 and 7 show the correlations between teachers' ranks across each of these weighting schemes. Overall, these results indicate that a teacher's relative rank is not very sensitive to the exact weighting approach selected (among the three we examine).

Next, we look more generally at the value of these summary measures relative to the simpler single-dimension value-added measures typically used in evaluations. First, we compute test and non-test value-added measures, in both the single and multi-dimensional framework. The non-test measures are averages of teacher effects on future grades, current attendance, and future attendance. 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. Columns (4) to (7) of Table 7 show the correlation between our summary measures of effectiveness and empirical Bayes estimates of effectiveness based on test and non-test score outcomes.

In general, the non-test empirical Bayes estimates are highly correlated with the summary measures of effectiveness. The test-based measures are only moderately correlated with the summary measures, and much less correlated than the non-test measures. For example, among elementary school teachers, ratings based on the regression weights have a correlation of 0.91 with single dimension non-test value-added and 0.33 for single dimension test value-added. The differences are slightly smaller for the other summary measures and middle school, but the patterns are similar. Figures 6 and 7 also show these correlations, and make it clear that the non-test value-added measures are very close to the summary measures, while test-based value-added is much less correlated.

Overall, these results indicate that some simple approaches, such as average measures based on non-test outcomes, perform quite well. However, they also highlight some weaknesses to just relying on test score value-added. Tables 8 and 9 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 5% 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. If the goal is to improve test scores, test score value-added will have the largest impact, and the first-eigenvalue summary measure preforms better than the other summary measures and the non-test measure. Decisions based on the summary measures are also projected to improve non-test outcomes more than those based on test score value-added.

For middle school (Table 9), the differences are smaller. Replacing the bottom 5% in terms of test score value-added is projected to increase high school graduation by 7.8pp relative to 6.7 to 6.9 pp for decisions based on the summary measures. For non-test score

outcomes, the expected gains from replacing the bottom 5% are about 6 times larger if decisions are not based on test score value-added.

Panel (B) in Tables 8 and 9 shows the overlap in teachers who are in the bottom 5% 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 5% 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 5% on test score value-added, only 16% are in the bottom 5% for non-test value-added and 37% are in the bottom 5% for the eigenvalue summary measure. Among middle school teachers, there is relatively high overlap (77-83%) in terms of the bottom 5% 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 incentives 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.

VI Teachers with Missing Outcomes

In order to present a clear picture of the multidimensionality of teacher effects, we have made some simplifications. For example, we only incorporate a single year into the empirical Bayes estimates; in practice, most evaluation policies incorporate measures from multiple years. Doing so, however, is relatively straightforward and Appendix F discusses this extension.We have also assumed that we observe noisy measures of teacher effectiveness for all of the outcomes we aim to predict. This would not be true if, for example, one wants to estimate effectiveness for teachers in both tested and non-tested subjects/grades, in which case teachers in non-tested subjects/grades would be missing test score measures. We now walk through how to estimate teacher effects when different teachers have different sets of observed measures. We first discuss how to construct empirical Bayes estimates when researchers do not observe noisy measures of all outcomes. Then we discuss the policy implications of having the set of available measures vary across teachers and how to ensure policies do not advantage or disadvantage workers based on which measures are observed.

VI.A Multidimensional Empirical Bayes Estimates with Missing Data

We can compute the posterior empirical Bayes distribution of a teacher's effect on some set of measures when we observe a different set using an approach that is equivalent to how, in section III, we used the estimates of a teacher's effect on her students' short-term outcomes to compute the empirical Bayes estimate of her effect on graduation. Below, we describe how to compute the multidimensional empirical Bayes estimates of a vector of outcomes when only a subset of noisy measures are observed.

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.[30] 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 $\Sigma_j = \begin{bmatrix} \Sigma_{1,1,j} & \Sigma_{1,2,j} \\ \Sigma_{2,1,j} & \Sigma_{2,2,j} \end{bmatrix}$.

The aim of this section is to compute how the true teacher effects are distributed conditional on the observed noisy measures, i.e., to compute the distribution of $\Theta_j|\theta_{2,j,t-1}$. Given our partitions, it is relatively straightforward to derive that:

$$\Theta_j|\theta_{2,j,t-1} \sim N\Big(\mu_j, V_j\Big) \tag{16}$$

where:

$$\mu_j = \begin{bmatrix} \Omega_{1,2} \\ \Omega_{2,2} \end{bmatrix} \left(\Omega_{2,2} + \Sigma_{j,2,2}\right)^{-1} \theta_{2,j,t-1} \qquad \text{and} \qquad V_j = CBC' + D$$

where

$$B = \left(\Omega_{2,2}^{-1} + \Sigma_{j,2,2}^{-1}\right)^{-1} \qquad C = \begin{bmatrix} \Omega_{1,2}\Omega_{2,2}^{-1} \\ I \end{bmatrix} \qquad D = \begin{bmatrix} \Omega_{1,1} - \Omega_{1,2}\Omega_{2,2}^{-1}\Omega_{2,1} & 0 \\ 0 & 0 \end{bmatrix}$$

and I is the identity matrix where the number of rows equals the number of observed

[30] 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.

measures.

We leave the derivation of this expression to Appendix F and focus here on the intuition behind the approach and the assumptions required. First, it is worth briefly contrasting this with a natural alternative, in which one first imputes values for the unobserved measures. After doing the imputation, no measures are missing and so one could use the main approach defined in Section III to compute the empirical Bayes estimates and/or to summarize teacher effectiveness. While intuitive, this imputation approach errs in that it treats the imputed values as additional data to condition on rather than estimates imputed from the other observations. As we show in Appendix G, under two natural imputation approaches this leads to too much shrinkage in the resulting empirical Bayes estimates.

Second, 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 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.

VI.B Fairness in Personnel Policy with Missing Data

We next turn our attention to the practical matter of how to use the empirical Bayes estimates from above to conduct teacher personnel policy in a context where not all measures are observed for all teachers. Importantly, we want to ensure that the policy does not advantage or disadvantage workers depending on which measures are observed.

For our exploration of this question, we take as given the ideal policy that a principal would choose if she observed the teachers' true effectiveness on all dimensions. This policy can be thought of as a function that maps a teacher's true effectiveness to a real number that corresponds to some decision. We will denote this function as $d(\Theta_j)$, with d representing the fact that it corresponds to a principal's decision. Examples include cases where the output is constrained to be a) either zero or one, depending on whether the teacher is retained or fired, b) one of a handful of values depending on the teacher's proficiency status, or c) the dollar amount of the bonus each teacher receives.

The challenge is that $d(\Theta_j)$ is not implementable since Θ_j is not directly observed. This is because the observed measures contain noise and because we only observe a subset of the measures for each individual, with the the subset of observed measures varying across individuals. As a possible solution, note that the approach defined in Section VI.A constructs the best estimate of the teachers' true effects on all of the the student outcomes, even the ones that researchers do not observe. A natural approach is therefore to use the resulting value-added measures in the function d. Formally, this would consist of the policy being $d(\hat{\Theta}_j)$, where $\hat{\Theta}_j = \begin{bmatrix} \Omega_{1,2} \\ \Omega_{2,2} \end{bmatrix} \left(\Omega_{2,2} + \Sigma_{j,2,2}\right)^{-1} \theta_{2,j,t-1}$.

For example, consider the case where the bottom 5% of teachers are removed from their jobs (which is a typical benchmark for policy exercises related to value-added and personnel policies).[31] If true effectiveness was observed, the policy would order teachers by $\omega'\Theta_j$ and replace teachers in the bottom 5% of this ordering, i.e., replacing the bottom 5% of teachers as ranked by their true effectiveness. Since Θ_j is not observed, the policy instead orders teachers by $\omega'\hat{\Theta}_j$ and replaces teachers in the bottom 5% of this ordering, i.e., replacing the bottom 5% of teachers as ranked by the empirical Bayes estimates of their effectiveness.

While intuitive, this policy is not fair, in the sense that it systematically advantages/ disadvantages individuals depending on which measures are observed. To see this empirically, we conducted a simulation in which we dropped each measure for each teacher, with

[31] Here we focus on summary measures based on the weights implied by the first eigenvalue, as discussed in Section IV.

25% probability. This led to variation across teachers in the number of measures observed that is independent from the teacher's true ability. However, the dark grey bars in Figure 8 show that the probability a teacher's value-added score, i.e., their $\omega'\hat{\Theta}_j$, is in the bottom 5% of the distribution depends in large part on how many measures were randomly dropped. In particular, the more measures that are observed the more likely it is that the teachers' value of $\omega'\hat{\Theta}_j$ is in the bottom 5% of the distribution.

The issue is that by substituting $\hat{\Theta}_j$ for Θ_j, the policy essentially treats the estimated value-added scores as the teachers' true effects. This means the policy ignores differences across teachers in the uncertainty of these estimates. In the context of empirical Bayes estimates, teacher effects with more uncertainty are more "shrunken" towards the overall mean (of zero). Since shrinking the estimates toward zero makes it less likely to be in the bottom 5%, this means the policy generally rewards teachers for whom we only observe a few measures, i.e., whose true effects are more uncertain, and punishes teachers for whom we observe more measures, i.e., whose true effects are more certain. Note that this issue also arises when multiple years of data are included in the evaluation and a similar result shows that teachers in the data for more years are more likely to fall in the bottom 5%.

Incorporating information about the posterior variance, in addition to the posterior mean, can help ensure the ranking of teachers, and the resulting policy decisions, are not affected by how many measures are observed. Specifically, we propose a new way to implement the ideal policy. Suppose that the policy can be defined by a function $d(\Theta_j)$, which maps the teachers' true effectiveness to a real number that reflects the policy.[32] For example, $d(\Theta_j)$ be defined as one if the teacher is retained and zero otherwise or correspond to the amount of bonus they receive. Instead of substituting $\hat{\Theta}_j$ into d, we propose computing the expected value of $d(\Theta_j)$ using the posterior distribution of Θ_j. That is, we define the

[32]Throughout, we will assume that $d(\Theta_j)$ a random variable, i.e., a well-behaved function for which we can compute its expectation.

new policy's value as:

$$\mathbb{E}_{\Theta_j}[d(\Theta_j)|\theta_{j,t-1}] \equiv \int d(\Theta_j) f(\Theta_j|\theta_{j,t-1}) d\Theta_j$$

where $f(\Theta_j|\theta_{j,t-1})$ is the posterior distribution of Θ_j conditional on $\theta_{j,t-1}$. The proposed policy is then a function of $\theta_{j,t-1}$, the observed measures, so we denote this policy as $\mathcal{P}(\theta)$, ignoring the $j, t-1$ subscripts to limit notation.[33]

Unlike the policy of $d(\hat{\Theta}_j)$, $\mathcal{P}(\theta)$ incorporates information about the entire posterior distribution of Θ_j, rather than just the posterior mean. This helps ensure that the policy does not discriminate towards individuals depending on which measures are observed. Formally, we have the following theorem, which follows from the law of iterated expectations:[34]

Theorem 1. *Let θ_0 to be the observed measures for some teacher and θ_1 to be any subset of the unobserved measures for the same teacher. Further, let $d(\Theta_j)$ be any policy that a principal would like to implement if Θ_j were fully observed and define $\mathcal{P}(\theta) = \mathbb{E}_{\Theta_j}[d(\Theta_j)|\theta]$. Then, for any θ_0:*

$$\mathcal{P}(\theta_0) = \mathbb{E}_{\theta_1}\Big[\mathcal{P}(\theta_1, \theta_0) \;\Big|\; \theta_0\Big]$$

Observing additional measures will generally impact the likelihood that a teacher is retained, i.e., $\mathcal{P}(\theta_1, \theta_0) \neq \mathcal{P}(\theta_0)$. The theorem says, however, that under this policy approach, the changes will not consistently skew either positive or negative, regardless of what measures are observed and their values. One implication of this fairness condition is that, absent teachers having private information about their own ability, every teacher is indifferent about whether more measures are added.[35]

Note also that Theorem 1 uses a strong definition of fairness. We could consider a

[33]Technically, it is also a function of the number of students taught by the teacher, since that also determines the posterior distribution. We will ignore that dependence in the function to keep the notation simple. Note also that the domain of this function is complex, since the number and type of observed measures in θ will differ across different individuals.

[34]The full proof is in Appendix B.

[35]More formally, this assumes that the teachers own beliefs about the θ_1 they would get, conditional on θ_0, is the same as the computed marginal distribution, i.e., $f(\theta_1|\theta_0)$. It also assumes that teachers are risk neutral regarding the values of d.

weaker definition of fairness that only requires that $\mathbb{E}_{\theta_1,\theta_0}\big[\mathcal{P}(\theta_1,\theta_0)\big] = \mathbb{E}_{\theta_0}\big[\mathcal{P}(\theta_0)\big]$ for all subsets of measures. This states that before any of the measures are observed, the subset of measures that will eventually be observed will not impact the expectation of the policy. The definition of fairness we use implies this condition, but not vice versa.

To see the theoretical result empirically, we return to the simulation example discussed above where the bottom 5% of teachers are removed from the school and we randomly removed measures from some teachers. In contrast to the dark grey bars, the light grey bars in Figure 8 show that the policy defined by $\mathcal{P}(\theta) = \mathbb{E}_{\Theta_j}[d(\Theta_j)|\theta]$ ensures that the probability a teacher is retained is the same regardless of how many measures are observed.

While the results in Figure 8 seem promising, there are a handful of caveats we want to emphasize. First, in the empirical example the policy results in a probability that each teacher is retained, rather than a binary answer of "retain" or "remove." Actually implementing such a probabilistic policy is unlikely and its implementation would be fraught with logistical, ethical, and political challenges. Second, the "fairness constraint," i.e., the requirement $\mathcal{P}(\theta_0) = \mathbb{E}_{\theta_1}\Big[\mathcal{P}(\theta_1,\theta_0) \;\Big|\; \theta_0\Big]$ for all θ_0, is not without cost. For example, the expected value-added increase is larger for a policy that removes the teachers in the bottom 5% of the measured value-added distribution than a policy that removes teachers with a likelihood proportional to the probability of their true effectiveness being in the bottom 5%. A further exploration of the trade-off between efficiency and fairness under a variety of objectives would be an interesting avenue of further research.

Finally, we note that our assumption that both the error terms and true effects are normally distributed is important here, in contrast to the assumption about the normality of the value-added estimates themselves.[36] When the policy aims to remove the bottom 5% of teachers, the value of $\mathbb{E}[d(\Theta_j)|\theta]$ depends greatly on the tails of the posterior distribution. If the parametric assumptions are wrong, the policy is no longer fair.[37]

[36]The value-added estimates can be thought of as the mean of the posterior distribution under the assumption of normality or as the best linear predictor of the teachers' true effects under weaker assumptions.

[37]This shows up in the fact that the expectation in $\mathbb{E}_{\theta_1}\big[\mathcal{P}](\theta_1,\theta_0)|\theta_0]$ is defined using the true distribution of θ_1 conditional on θ_0, while the expectation in $\mathbb{E}_{\Theta_j}[d(\Theta_j)|\theta]$ uses the estimated posterior distribution.

VII 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 components analysis and using the empirical Bayes estimates as covariates. The methods used to compute empirical Bayes estimates also influences 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 summary measures are only moderately correlated with traditional test score value-added, and there is little overlap between teachers who are at the bottom 5% in terms of the summary measures and test score value-added.

We conclude with a discussion of how to use one of the summary measures when not all components of the measure are observed for all teachers. We show that, as long as the full variance/covariance matrix of the teacher's true effects can be estimated, the multidimensional empirical Bayes approach can easily be extended to cases where not all measures are observed. However, we illustrate that using the resulting value-added measures without

accounting for their uncertainty advantages or disadvantages individuals depending on how many of their measures are observed. We then show that using information from the full posterior distribution ensures that the resulting policy is fair.

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.

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.

Aucejo, Esteban M., Patrick Coate, Jane Cooley Fruehwirth, Sean Kelly, and Zachary Mozenter, "Match Effects in the Teacher Labor Market: Teacher Effectiveness and Classroom Composition," 2020.

Aucejo, Esteban, Teresa Romano, and Eric S. Taylor, "Does evaluation change teacher effort and performance? Quasi-experimental evidence from a policy of retesting students," *Review of Economics and Statistics*, Forthcoming.

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.

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.

Dee, Thomas S. and James Wyckoff, "Incentives, Selection, and Teacher Performance: Evidence from IMPACT," *Journal of Policy Analysis and Management*, 2015, *34* (2), 267–297.

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

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.

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.

Macartney, Hugh, "The Dynamic Effects of Educational Accountability," *Journal of Labor Economics*, 2016, *34* (1), 1–28.

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.

Papay, John P., Eric S. Taylor, John H. Tyler, and Mary E. Laski, "Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data," *American Economic Journal: Economic Policy*, 2020, *12* (1), 359–88.

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.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor, "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools," *American Economic Review*, 2012, *102* (7), 3184–3213.

Taylor, Eric S. and John H. Tyler, "The Effect of Evaluation on Teacher Performance," *American Economic Review*, 2012, *102* (7), 3628–3651.

VIII 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	7061	0	13912	0
Teacher-Years	20683	0	48617	0
Teacher-Subject-Years	20868	0	52312	0
Students	183165	0	617563	0
Student-Years	477286	0	1482360	0
Student-Subject-Years	477286	0	2692055	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	0.205	0.144	0.294
ELA Test	0.162	0.107	0.260
Future Math Test	0.188	0.129	0.263
Future ELA Test	0.153	0.101	0.240
Attendance	0.011	0.005	0.037
Future Attendance	0.122	0.074	0.218
Future Grades Math	0.200	0.120	0.383
Future ELA Grades	0.199	0.109	0.345
(B) Middle School			
Test Scores	0.165	0.117	0.232
Future Test Scores	0.177	0.122	0.308
Attendance	0.013	0.008	0.033
Future Attendance	0.164	0.097	0.299
Future Grades in Subject	0.263	0.144	0.472
Future Grades Other Subjects	0.348	0.223	0.514

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)	Future Math Test (3)	Future ELA Test (4)	Attendance (5)	Future Attendance (6)	Future Math Grades (7)	Future ELA Grades (8)
Math Test	1.000	0.715	0.392	0.333	0.061	0.065	0.085	0.168
ELA Test	0.715	1.000	0.339	0.549	0.039	0.195	0.091	0.198
Future Math Test	0.392	0.339	1.000	0.809	-0.132	0.357	0.388	0.460
Future ELA Test	0.333	0.549	0.809	1.000	-0.126	0.364	0.405	0.439
Attendance	0.061	0.039	-0.132	-0.126	1.000	0.085	-0.061	-0.059
Future Attendance	0.065	0.195	0.357	0.364	0.085	1.000	0.305	0.316
Future Grades Math	0.085	0.091	0.388	0.405	-0.061	0.305	1.000	0.816
Future ELA Grades	0.168	0.198	0.460	0.439	-0.059	0.316	0.816	1.000

	(B) Middle School					
	Test Scores (1)	Future Test Scores (2)	Attendance (3)	Future Attendance (4)	Future Grades in Subject (5)	Future Grades in Other Subjects (6)
Test Scores	1.000	0.887	0.051	0.258	0.258	0.313
Future Test Scores	0.887	1.000	0.117	0.379	0.370	0.459
Attendance	0.051	0.117	1.000	0.397	0.142	0.201
Future Attendance	0.258	0.379	0.397	1.000	0.322	0.397
Future Grades in Subject	0.258	0.370	0.142	0.322	1.000	0.779
Future Grades Other Subjects	0.313	0.459	0.201	0.397	0.779	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	0.346	0.627	-0.428	-0.291
ELA Test	0.283	0.443	-0.172	0.513
Future Math Test	0.454	0.089	0.612	-0.404
Future ELA Test	0.369	0.092	0.442	0.137
Attendance	-0.002	0.003	-0.009	0.012
Future Attendance	0.161	-0.063	0.199	0.685
Future Grades Math	0.446	-0.482	-0.302	-0.029
Future ELA Grades	0.482	-0.397	-0.294	-0.019
(B) Middle School				
Test Scores	0.170	0.656	-0.235	-0.009
Future Test Scores	0.239	0.649	-0.122	-0.011
Attendance	0.007	0.002	0.031	0.009
Future Attendance	0.179	0.190	0.869	0.411
Future Grades in Subject	0.534	-0.262	-0.386	0.705
Future Grades Other Subjects	0.772	-0.208	0.155	-0.578

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: 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	0.026***		0.021***	
	(0.003)		(0.001)	
Second Component	-0.013***		0.005***	
	(0.003)		(0.001)	
Third Component	0.001		0.000	
	(0.003)		(0.001)	
Fourth Component	0.004*		-0.002**	
	(0.002)		(0.001)	
(B) Individual Measures				
Math Test		0.003		
		(0.007)		
ELA Test		0.000		
		(0.008)		
Test Score				0.002
				(0.006)
Future Math Test		0.006		
		(0.009)		
Future ELA Test		-0.002		
		(0.010)		
Future Test Score				0.006
				(0.007)
Attendance		-0.006*		-0.008***
		(0.004)		(0.001)
Future Attendance		0.012***		0.005***
		(0.004)		(0.002)
Future Grade		0.017*		-0.001
		(0.009)		(0.003)
Future Grade Other Subjects		0.001		0.018***
		(0.009)		(0.003)
N	2,918	2,918	14,128	14,128
Fraction Variance Explained	0.2355	0.2743	0.4138	0.4786

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. The last row reports the fraction of variance in teacher's effectiveness at increasing their students' high school graduation that is explained by the set of variables in each regression.

Table 6: 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
Future Math Test	33.044	50.465	23.878	28.835	40.864	23.176
Future ELA Test	22.429	33.027	14.717	15.905	21.732	11.608
Attendance	-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 Grades in Subject	6.114	11.733	27.438	5.682	10.118	28.362
Future Grades Other Subjects	20.015	22.476	15.067	18.538	19.318	15.523
(B) Middle School						
Test Scores	15.491	22.044	15.303	15.758	24.774	15.936
Future Test Scores	4.040	6.872	5.805	4.406	8.279	6.480
Attendance	40.109	40.808	40.810	3.314	3.725	3.452
Future Attendance	5.765	6.340	8.448	5.826	7.079	8.741
Future Grades in Subject	6.301	0.693	-0.993	10.199	1.238	-1.645
Future Grades Other Subjects	28.294	23.243	30.626	60.498	54.905	67.036

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 7: 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	0.943	0.904	0.619	0.612	0.802	0.731
PCA Regression	0.943	1.000	0.944	0.341	0.334	0.875	0.771
Regression	0.904	0.944	1.000	0.303	0.332	0.963	0.913
Multidim Test VA	0.619	0.341	0.303	1.000	0.967	0.143	0.144
Single Dim Test VA	0.612	0.334	0.332	0.967	1.000	0.189	0.203
Multidim Non-Test VA	0.802	0.875	0.963	0.143	0.189	1.000	0.970
Single Dim Non-Test VA	0.731	0.771	0.913	0.144	0.203	0.970	1.000
(B) Middle School							
PCA First Eigenvalue	1.000	0.979	0.989	0.519	0.435	0.982	0.925
PCA Regression	0.979	1.000	0.987	0.664	0.575	0.943	0.853
Regression	0.989	0.987	1.000	0.544	0.447	0.969	0.882
Multidimensional Test VA	0.519	0.664	0.544	1.000	0.940	0.417	0.305
Single Dimension Test VA	0.435	0.575	0.447	0.940	1.000	0.327	0.222
Multidimensional Non-Test VA	0.982	0.943	0.969	0.417	0.327	1.000	0.960
Single Dimension Non-Test VA	0.925	0.853	0.882	0.305	0.222	0.960	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 in subject and future grades in other subjects. 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 8: 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

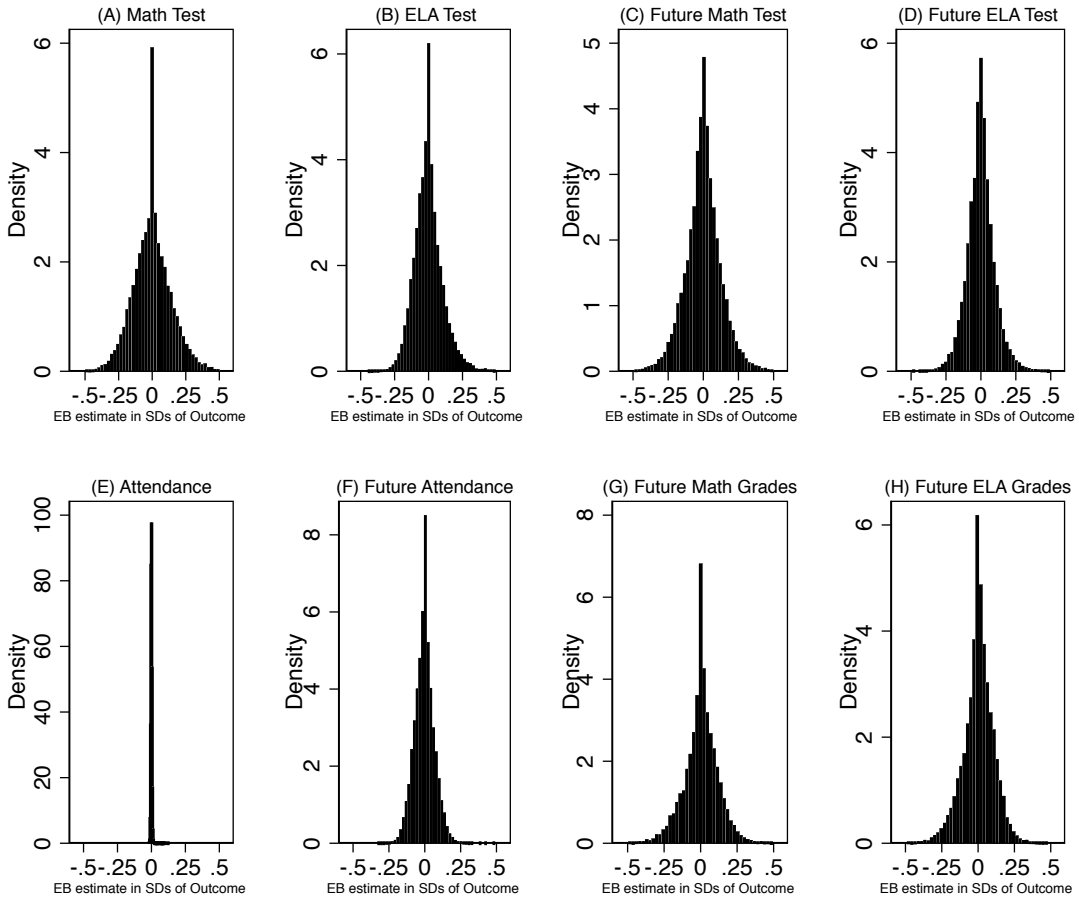
Notes: The estimates in Panel (A) show the differences in projected outcomes (high school graduation, test scores and non-test outcomes) for average teachers and those in the bottom 5% as ranked in terms of the value-added measure from the relevant column. The estimates in Panel (B) show the fraction 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 9: 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	0.067	0.069	0.067	0.078	0.064
Test Scores	0.252	0.322	0.250	0.968	0.159
Non-Test Outcomes	1.550	1.507	1.524	0.230	1.611
(B) Percent of Bottom 5% on Column VA also in Bottom 5% on Row VA					
HS Graduation	0.101	0.104	0.103	0.153	0.097
Test Scores	0.129	0.169	0.131	1.000	0.091
Non-Test Outcomes	0.827	0.772	0.792	0.091	1.000
Eigenvalue Summary	1.000	0.913	0.931	0.129	0.827

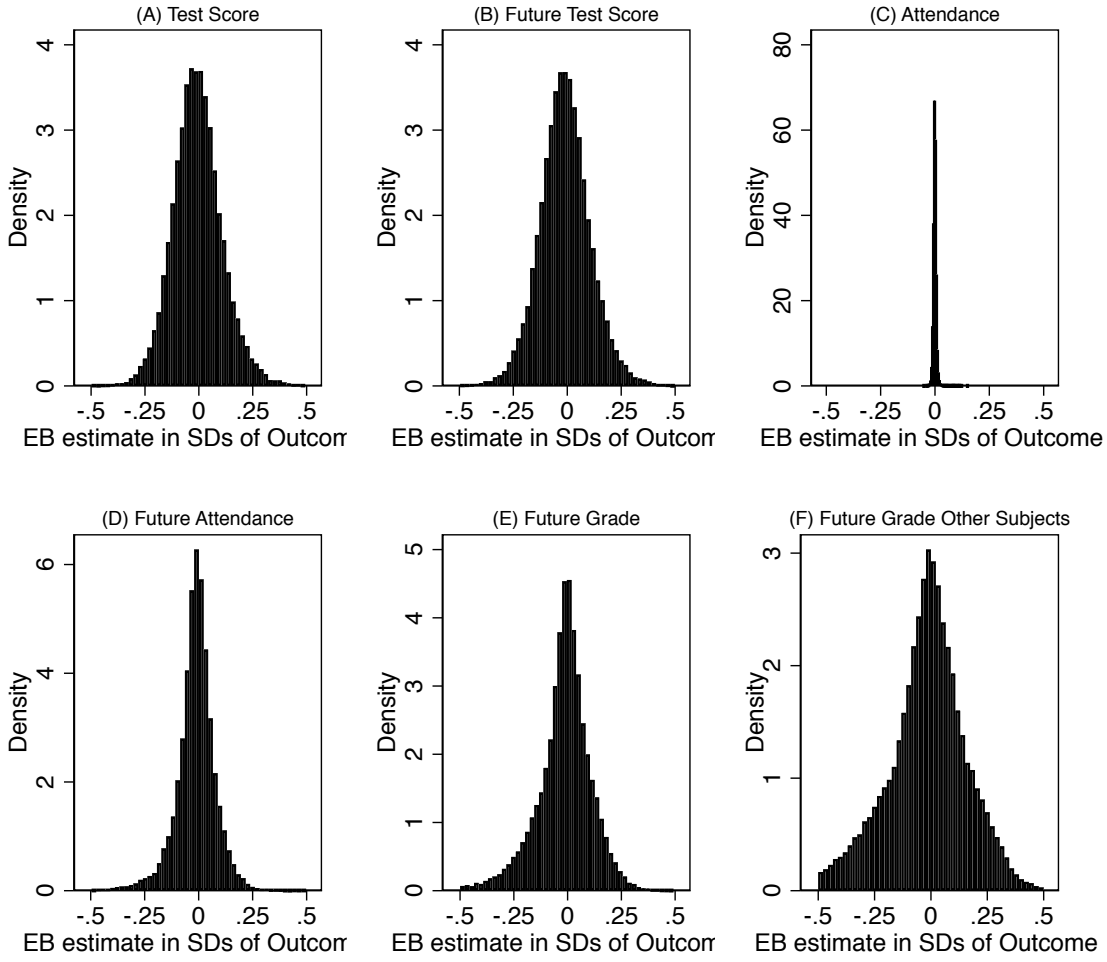
Notes: The estimates in Panel (A) show the differences in projected outcomes (high school graduation, test scores and non-test outcomes) for average teachers and those in the bottom 5% as ranked in terms of the value-added measure from the relevant column. The estimates in Panel (B) show the fraction 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, 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 middle school teachers (grades 6-7) and test score measures are based on the subject a teacher teaches.

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



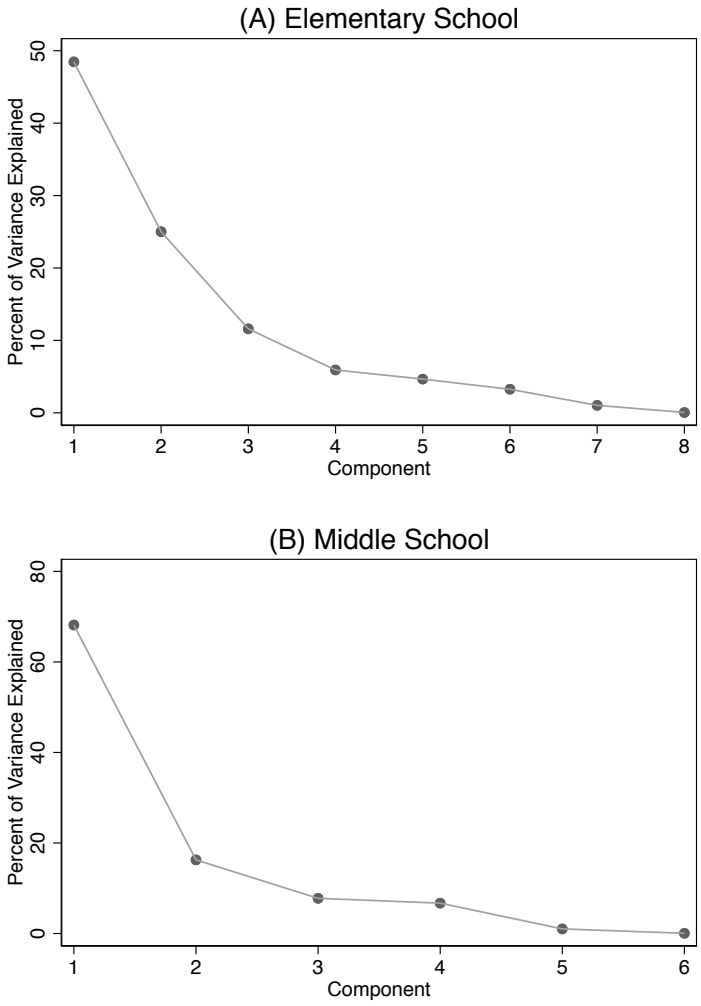
Notes: The figures above show the distribution of the multidimensional empirical Bayes estimates of teacher effects on individual outcomes for elementary school (5th grade). All estimates are in standard deviations of the outcome measure (standardized at the student level before computing teacher effects). 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



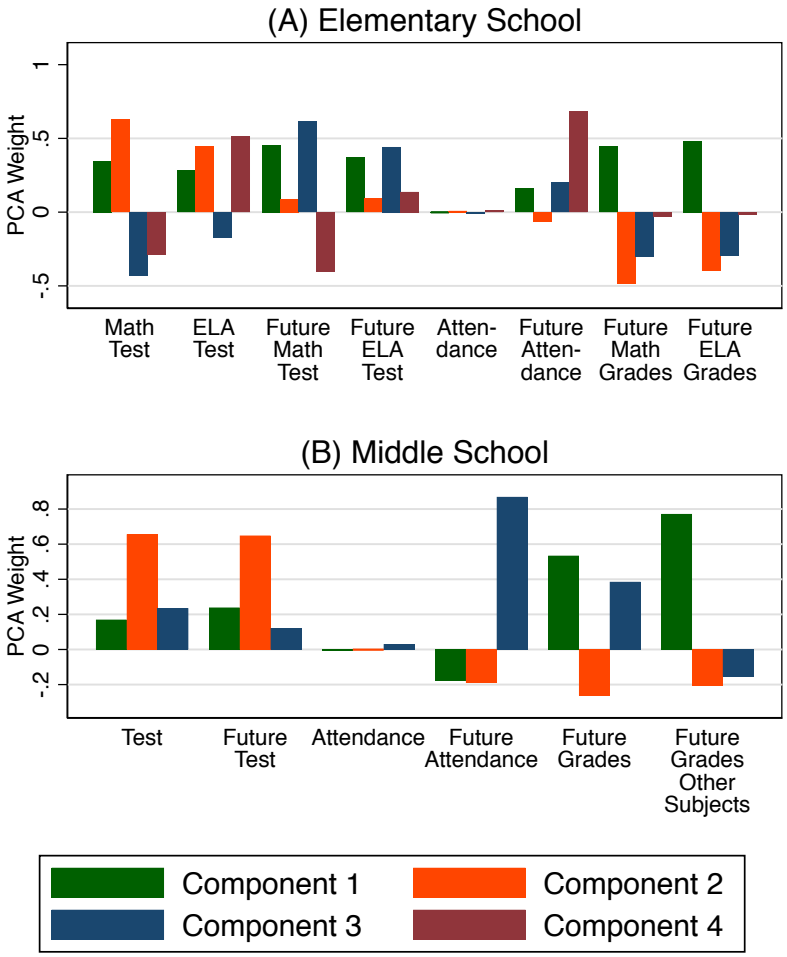
Notes: The figures above show the distribution of the multidimensional empirical Bayes estimates of teacher effects on individual outcomes for middle school (6th - 7th grade). All estimates are in standard deviations of the outcome measure (standardized at the student level before computing teacher effects). Panels A 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



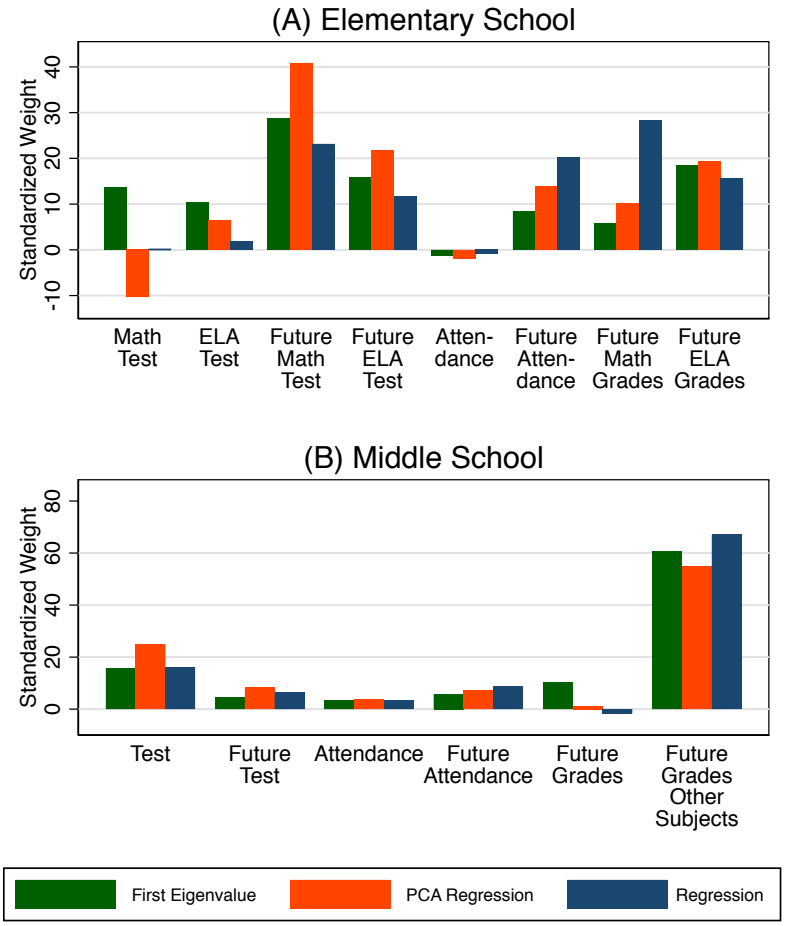
Notes: The figures above show the percent of variance in teacher effects on our student outcome measures explained by each principal component. These estimates come from conducting principal components analysis on the true measures of teacher effects. Panel (A) is for elementary school and is based on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

Figure 4: PCA Components



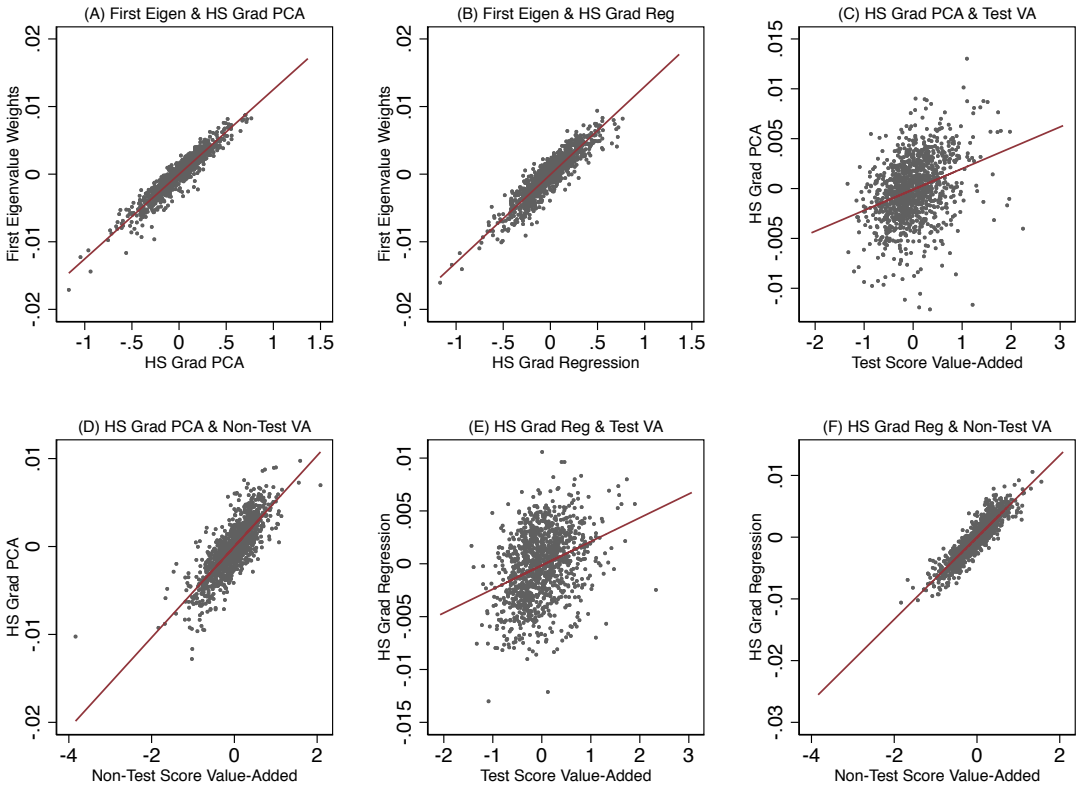
Notes: The figures above show the relative weight each student outcome receives in each of the four main principal components. For middle school (in panel B) test scores and future grades refer to the subject taught by the focal teacher. In elementary school (panel A) teachers teach both math and ELA. 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, 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



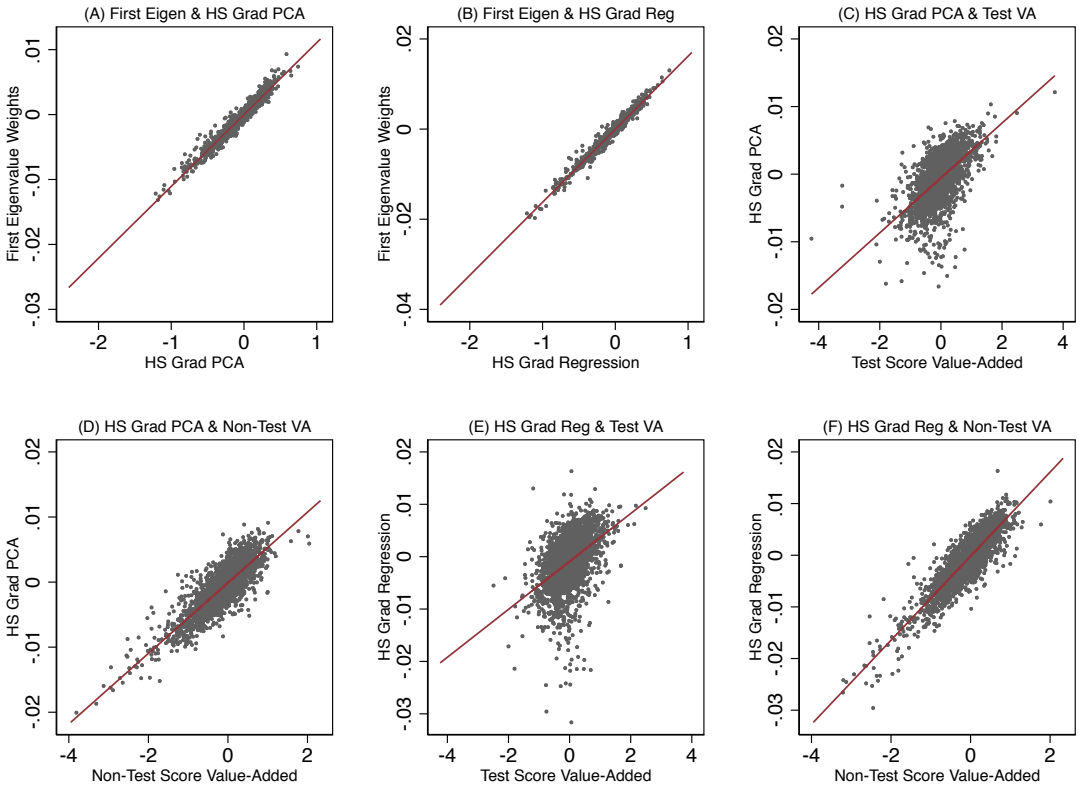
Notes: The figures above show the relative weights placed on each individual outcome for each of our three main approaches for creating summary measures of teacher effectiveness. The first approach (represented by the green bars) uses the first eigenvalue from principal components analysis to combine 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.

Figure 6: Elementary School: Correlation between Teacher Ratings on Different Measures of Effectiveness



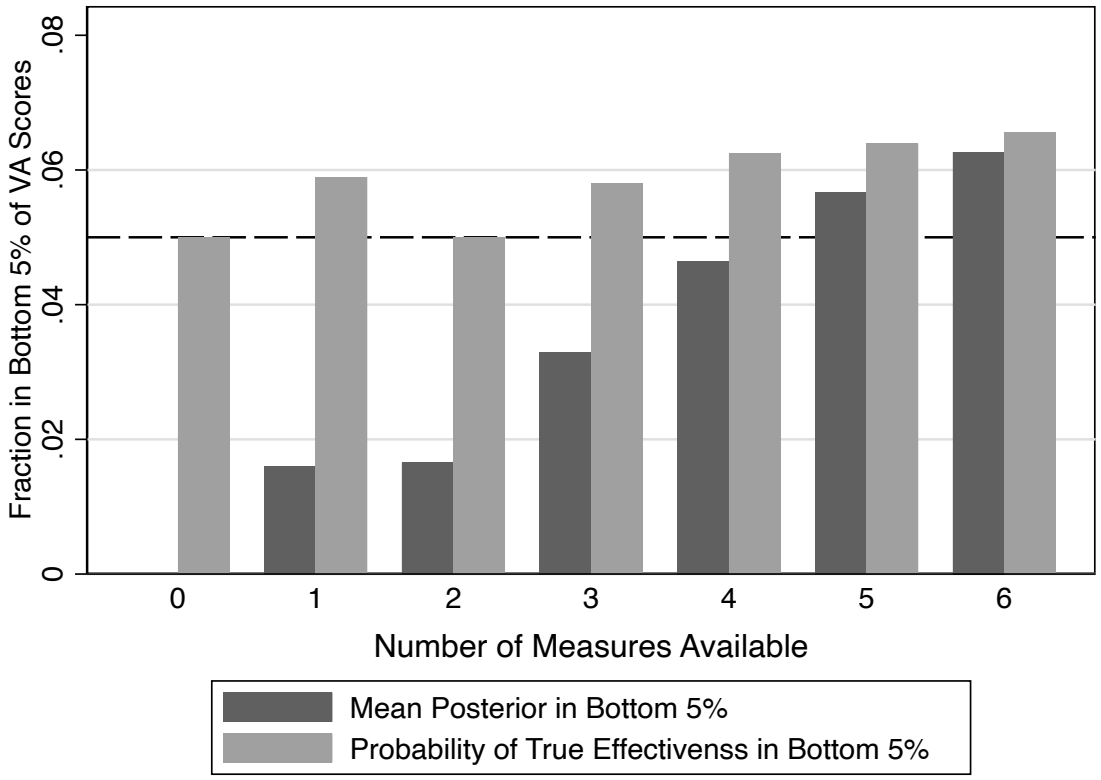
Notes: These figures show the correlations between elementary school teachers' ratings on our summary measures of effectiveness and traditional single dimensions value-added measures based on test scores or non-test score measures. In panels (A) and (B), the y-axis is based on the weights from the *first eigenvalue* from principal components 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 7: 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 components 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 8: Fairness in Policy with Missing Measures



Notes: This figure considers two policies: one of which removes teachers whose value-added is in the bottom 5% of the distribution (illustrated by the dark grey bars) and the other removes teachers with a probability equal to the probability that their true effectiveness is in the bottom 5% of the distribution (illustrated by the light grey bars). We then randomly remove measures from teachers to see how the probability that a teacher is removed varies with the number of measures observed.

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	0.485	0.526	0.428
Component 2	0.250	0.300	0.199
Component 3	0.116	0.103	0.104
Component 4	0.059	0.039	0.086
Component 5	0.047	0.023	0.072
Component 6	0.033	0.007	0.067
Component 7	0.010	0.002	0.041
Component 8	0.001	0.000	0.002
(B) Middle School			
Component 1	0.682	0.758	0.582
Component 2	0.163	0.166	0.144
Component 3	0.078	0.053	0.126
Component 4	0.067	0.020	0.098
Component 5	0.010	0.002	0.048
Component 6	0.001	0.000	0.001

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: Correlation between Multidimensional and Single Dimension Empirical Bayes Estimates

(A) Elementary School	Math Test Scores	ELA Test Scores	Future Math Test	Future ELA Test	Attendance	Future Attendance	Future Math Grades	Future ELA Grades
Correlation	0.964	0.900	0.958	0.899	0.873	0.931	0.962	0.941

(B) Middle School	Test Scores	Future Test Scores	Attendance	Future Attendance	Future Grades in Subject	Future Grades in Other Subjects
Correlation	0.943	0.816	0.850	0.931	0.898	0.983

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.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	0.027***	0.030***	0.020***	0.016***
	(0.002)	(0.002)	(0.001)	(0.001)
Second Component	-0.011***	-0.007***	0.006***	0.015***
	(0.003)	(0.002)	(0.001)	(0.001)
Third Component	0.003	0.009***	0.000	-0.002
	(0.003)	(0.002)	(0.001)	(0.001)
Fourth Component	0.003	0.002	-0.002	-0.002**
	(0.002)	(0.002)	(0.001)	(0.001)
(B) Individual Measures				
Math Test Score	0.005	0.010		
	(0.007)	(0.007)		
ELA Test Score	-0.003	-0.005		
	(0.008)	(0.008)		
Test Score			0.003	0.013**
			(0.006)	(0.005)
Future Math Test	0.005	0.010		
	(0.010)	(0.009)		
Future ELA Test	0.002	0.007		
	(0.010)	(0.010)		
Future Test Score			0.005	0.007
			(0.007)	(0.005)
Attendance	-0.008**	-0.011***	-0.009***	-0.001
	(0.003)	(0.004)	(0.001)	(0.002)
Future Attendance	0.013***	0.014***	0.007***	0.000
	(0.004)	(0.003)	(0.002)	(0.001)
Future Grade	0.012	0.019***	-0.001	-0.003
	(0.009)	(0.006)	(0.003)	(0.002)
Future Grade Other Subjects	0.004	-0.008	0.016***	0.010***
	(0.008)	(0.006)	(0.003)	(0.002)
N	2,918	2,918	14,128	14,128

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 diplomas 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.4: 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	0.936	0.898	0.599	0.600	0.799	0.736
PCA Regression	0.936	1.000	0.942	0.331	0.334	0.878	0.781
Regression	0.898	0.942	1.000	0.299	0.335	0.964	0.917
Multidim Test VA	0.599	0.331	0.299	1.000	0.965	0.150	0.155
Single Dim Test VA	0.600	0.334	0.335	0.965	1.000	0.198	0.212
Multidim Non-Test VA	0.799	0.878	0.964	0.150	0.198	1.000	0.973
Single Dim Non-Test VA	0.736	0.781	0.917	0.155	0.212	0.973	1.000
(B) Middle School							
PCA First Eigenvalue	1.000	0.977	0.986	0.535	0.466	0.980	0.909
PCA Regression	0.977	1.000	0.986	0.672	0.597	0.936	0.828
Regression	0.986	0.986	1.000	0.560	0.478	0.963	0.858
Multidim Test VA	0.535	0.672	0.560	1.000	0.952	0.434	0.317
Single Dim Test VA	0.466	0.597	0.478	0.952	1.000	0.363	0.249
Multidim Non-Test VA	0.980	0.936	0.963	0.434	0.363	1.000	0.953
Single Dim Non-Test VA	0.909	0.828	0.858	0.317	0.249	0.953	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 in subject and future grades in other subjects. 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 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
Future Math Test	52.099	26.899	55.594	37.070	42.581	26.235	46.348	34.942
Future ELA Test	32.519	15.808	31.887	22.697	21.599	12.529	21.603	17.386
Attendance	-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 Grades in Subject	7.550	18.041	-0.327	15.795	6.571	18.739	-0.291	15.856
Future Grades Other Subjects	20.751	16.520	17.298	3.532	18.002	17.102	15.307	3.534
(B) Middle School								
Test Scores	22.916	15.836	33.072	40.670	25.774	16.949	36.081	38.159
Future Test Scores	7.283	5.806	11.894	15.769	8.780	6.661	13.910	15.860
Attendance	40.066	40.827	30.573	19.286	3.660	3.549	2.709	1.470
Future Attendance	6.561	10.217	7.774	5.198	7.331	10.864	8.426	4.846
Future Grades in Subject	0.574	-0.967	-1.080	-4.257	1.026	-1.646	-1.875	-6.355
Future Grades Other Subjects	22.600	28.281	17.768	23.334	53.429	63.623	40.748	46.020

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: Middle School: Implications of Changing Evaluation Measures

	Weighted Summary Measures					Empirical Bayes Estimates	
	PCA First Eigenvalue (1)	HS Grad PCA Reg (2)	HS Grad Reg (3)	Regents Reg (4)	Advanced Regents Reg (5)	Test Value-Added (6)	Non-Test Value-Added (7)
(A) Projected Change in Outcomes from Replacing Bottom 5% with Mean Teacher							
HS Graduation	0.067	0.069	0.067	0.067	0.083	0.078	0.064
Regents Diploma	0.070	0.072	0.070	0.070	0.085	0.077	0.068
Advanced Regents	0.022	0.024	0.023	0.022	0.034	0.042	0.021
(B) Percent of Bottom 5% on Column VA also in Bottom 5% on Row VA							
HS Graduation	0.101	0.104	0.103	0.105	0.132	0.153	0.097
Regents Diploma	0.114	0.117	0.118	0.118	0.140	0.142	0.114
Advanced Regents	0.038	0.042	0.036	0.032	0.069	0.105	0.041

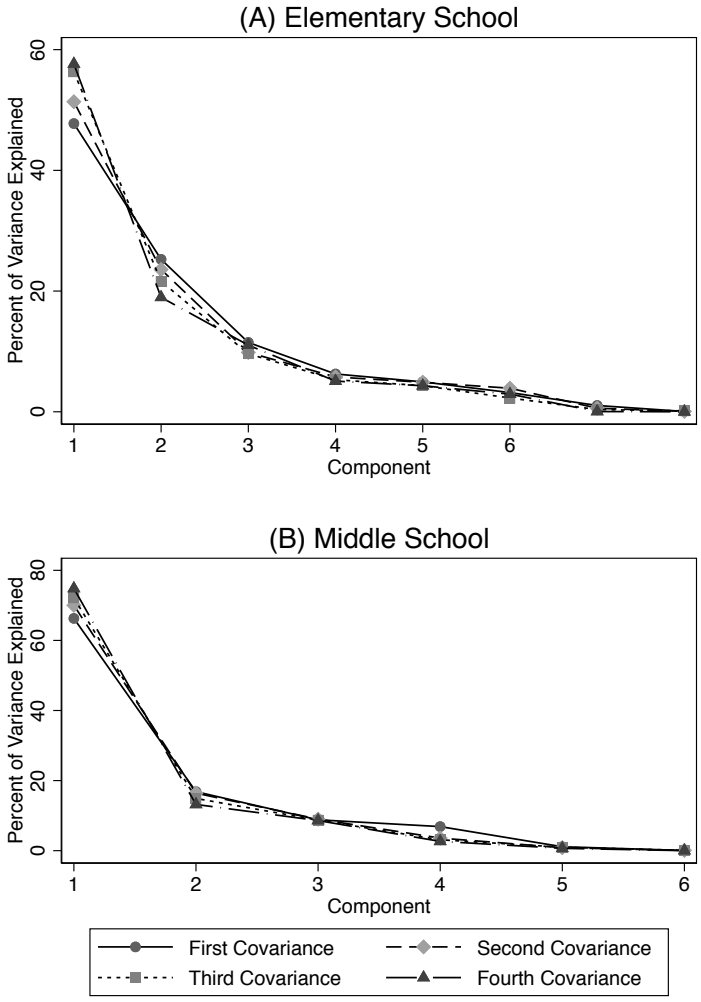
Notes: The estimates in Panel (A) show the differences in projected outcomes (high school graduation, test scores and non-test outcomes) for average teachers and those in the bottom 5% as ranked in terms of the value-added measure from the relevant column. The estimates in Panel (B) show the fraction 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 weights from a regression of teacher effects on receipt of a Regents diploma on the empirical Bayes estimates of teacher effects on individual outcomes. Column 5 is based on weights from a regression of teacher effects on earning an Advanced Regents diploma on the empirical Bayes estimates of teacher effects on individual outcomes. Column 6 is based on traditional estimates of teacher effects on test scores in the single dimension 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 in subject and future grades in other subjects. This measure equally weights teacher effects on these four outcomes. This table is based on middle school teachers (grades 6 and 7) and test score measures are based on the subject a teacher teaches.

Table A.7: Elementary School: Implications of Changing Evaluation Measures

	Weighted Summary Measures					Empirical Bayes Estimates	
	PCA First Eigenvalue	HS Grad PCA Reg	HS Grad Reg	Regents Reg	Advanced Regents Reg	Test Value-Added	Non-Test Value-Added
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(A) Projected Change in Outcomes from Replacing Bottom 5% with Mean Teacher							
HS Graduation	0.149	0.145	0.142	0.151	0.145	0.078	0.111
Regents Diploma	0.165	0.159	0.158	0.167	0.159	0.082	0.121
Advanced Regents	0.091	0.090	0.090	0.094	0.094	0.054	0.062
(B) Percent of Bottom 5% on Column VA also in Bottom 5% on Row VA							
HS Graduation	0.234	0.192	0.215	0.224	0.196	0.168	0.178
Regents Diploma	0.262	0.238	0.257	0.271	0.252	0.187	0.187
Advanced Regents	0.173	0.164	0.164	0.164	0.173	0.093	0.093

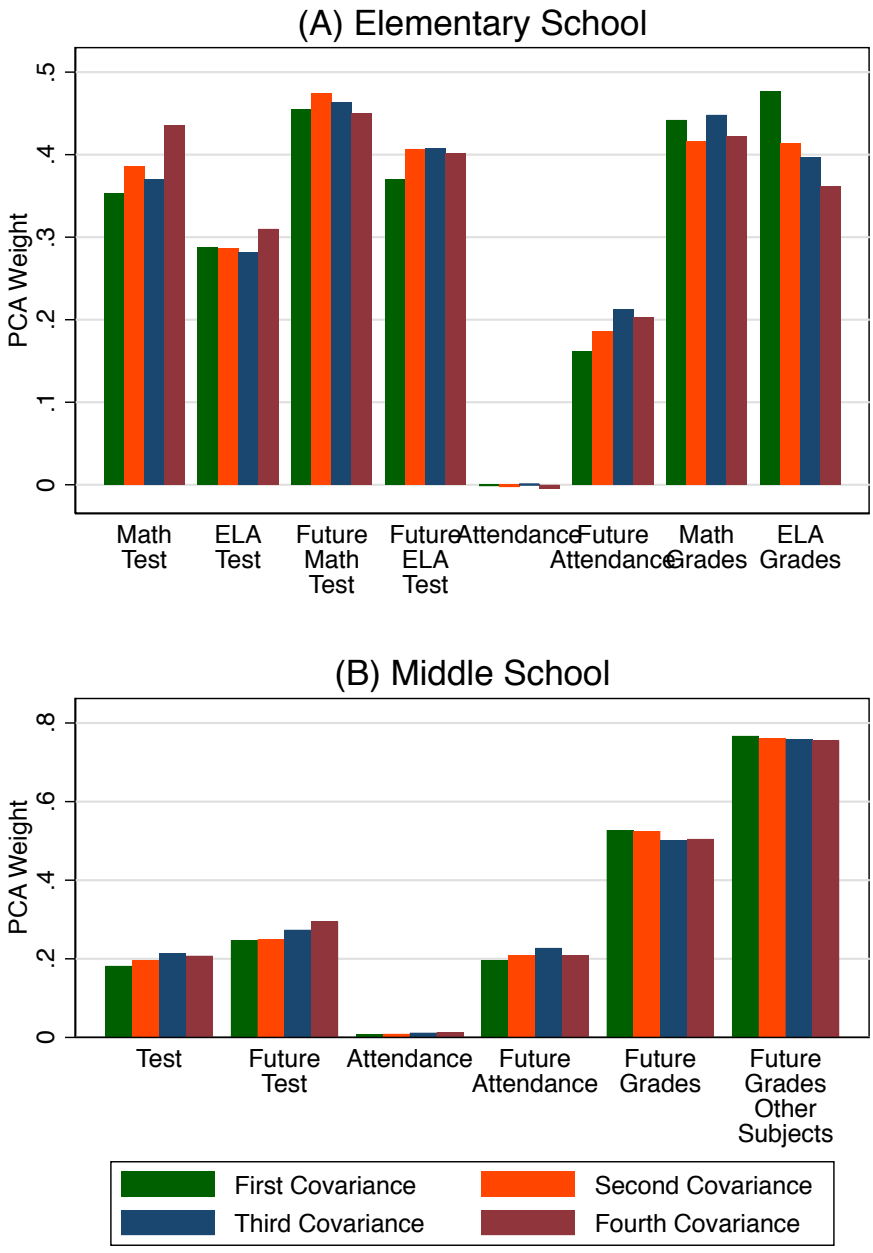
Notes: The estimates in Panel (A) show the differences in projected outcomes (high school graduation, test scores and non-test outcomes) for average teachers and those in the bottom 5% as ranked in terms of the value-added measure from the relevant column. The estimates in Panel (B) show the fraction 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 weights from a regression of teacher effects on receipt of a Regents diploma on the empirical Bayes estimates of teacher effects on individual outcomes. Column 5 is based on weights from a regression of teacher effects on earning an Advanced Regents diploma on the empirical Bayes estimates of teacher effects on individual outcomes. Column 6 is based on traditional estimates of teacher effects on test scores in the single dimension 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 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 measures are based on averages across math and reading.

Figure A.1: Scree Plot of Eigenvalues



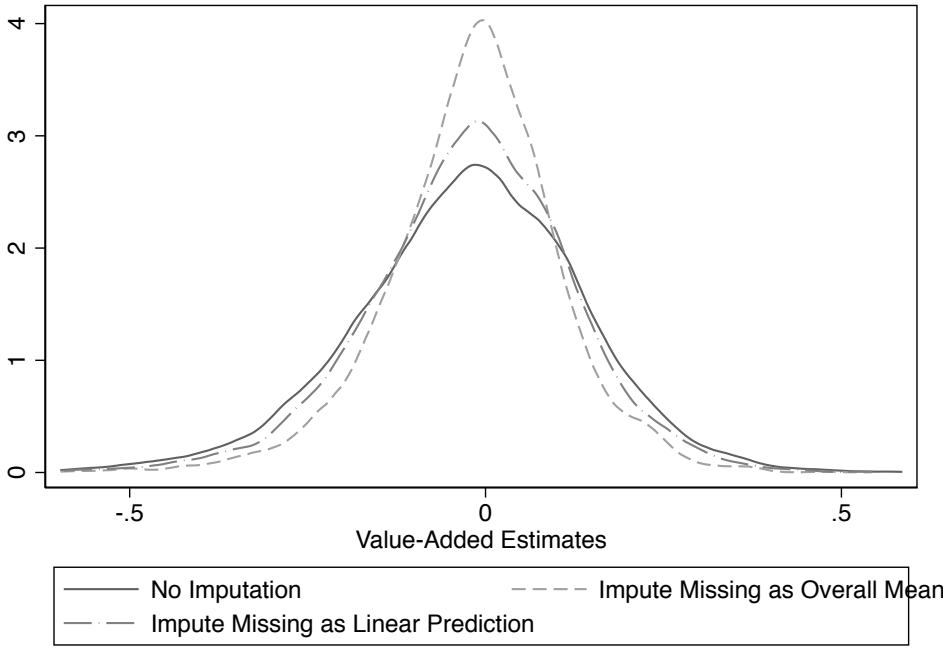
Notes: The figures above show the percent of variance in teacher effects on our student outcome measures explained by each principal component. These estimates come from conducting principal components analysis on four covariance matrices: $\mathbb{C}_1$, $\mathbb{C}_2$, $\mathbb{C}_3$, $\mathbb{C}_4$, 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.2: 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.3: 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 III. They differ in how the missing observations are handled. The "No Imputation" method uses the approach defined in Section VI. 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 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:*

$$\omega_{1,1} = \frac{1}{det(\Omega + \Sigma_j)}\left[\sigma^2_{\Omega,1}\sigma^2_{\Omega,2} + \sigma^2_{\Omega,1}\sigma^2_{\Sigma,2} - \rho^2_\Omega - \rho_\Omega\rho_\Sigma\right] \tag{17}$$

$$\omega_{1,2} = \frac{1}{det(\Omega + \Sigma_j)}\left[\sigma^2_{\Sigma,1}\rho_\Omega - \sigma^2_{\Omega,1}\rho_\Sigma\right] \tag{18}$$

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:

$$\Omega^*_j = \left((\Omega + \Sigma_j)^{-1}\Omega\right)' \tag{19}$$

$$= \left[\frac{1}{det(\Omega + \Sigma_j)}\begin{pmatrix} \sigma^2_{\Omega,2} + \sigma^2_{\Sigma,2} & -(\rho_\Omega + \rho_\Sigma) \\ -(\rho_\Omega + \rho_\Sigma) & \sigma^2_{\Omega,1} + \sigma^2_{\Sigma,1} \end{pmatrix}\begin{pmatrix} \sigma^2_{\Omega,1} & \rho_\Omega \\ \rho_\Omega & \sigma^2_{\Omega,2} \end{pmatrix}\right]' \tag{20}$$

$$\tag{21}$$

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

$$\omega_{1,1} = \frac{1}{det(\Omega + \Sigma_j)}\Big[(\sigma^2_{\Omega,2} + \sigma^2_{\Sigma,2})\sigma^2_{\Omega,1} - \rho_\Omega(\rho_\Omega + \rho_\Sigma)\Big] \tag{22}$$

$$\omega_{1,2} = \frac{1}{det(\Omega + \Sigma_j)}\Big[(\sigma^2_{\Omega,2} + \sigma^2_{\Sigma,2})\rho_\Omega - \sigma^2_{\Omega,2}(\rho_\Omega + \rho_\Sigma)\Big] \tag{23}$$

$$= \frac{1}{det(\Omega + \Sigma_j)}\Big[\sigma^2_{\Sigma,2}\rho_\Omega - \sigma^2_{\Omega,2}\rho_\Sigma\Big] \tag{24}$$

$$\omega_{2,1} = \frac{1}{det(\Omega + \Sigma_j)}\Big[(\sigma^2_{\Omega,1} + \sigma^2_{\Sigma,1})\rho_\Omega - \sigma^2_{\Sigma,1}(\rho_\Omega + \rho_\Sigma)\Big] \tag{25}$$

$$= \frac{1}{det(\Omega + \Sigma_j)}\Big[\sigma^2_{\Sigma,1}\rho_\Omega - \sigma^2_{\Omega,1}\rho_\Sigma\Big] \tag{26}$$

$$\omega_{2,2} = \frac{1}{det(\Omega + \Sigma_j)}\Big[(\sigma^2_{\Omega,1} + \sigma^2_{\Sigma,1})\sigma^2_{\Omega,2} - \rho_\Omega(\rho_\Omega + \rho_\Sigma)\Big] \tag{27}$$

□

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

$$(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1} = \left((\Omega + \Sigma_j)^{-1}\Omega\right)' \tag{28}$$

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:

$$\left[(\Sigma_j^{-1} + \Omega^{-1})^{-1}\Sigma_j^{-1}\right]^{-1} = \Sigma_j(\Sigma_j^{-1} + \Omega^{-1}) \tag{29}$$

$$= \Sigma_j\Sigma_j^{-1} + \Sigma_j\Omega^{-1} \tag{30}$$

$$= \mathbf{I} + \Sigma_j\Omega^{-1} \tag{31}$$

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:

$$\left[\left((\Omega + \Sigma_j)^{-1}\Omega\right)'\right]^{-1} = \left[\Omega(\Omega + \Sigma_j)^{-1}\right]^{-1} \tag{32}$$

$$= (\Omega + \Sigma_j)\Omega^{-1} \tag{33}$$

$$= \mathbf{I} + \Sigma_j\Omega^{-1} \tag{34}$$

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 III.A, the model specified in Section III.C, and the assumption that the number of teachers and students increases to infinity, we have that:* $\hat{\beta} \to \beta$, $\hat{\Omega} \to \Omega$, $\hat{\Sigma}_\epsilon \to \Sigma_\epsilon$ *and* $\hat{\Sigma}_\nu \to \Sigma_\nu$.

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 $\overline{X}_{j,t-1}$ as the average outcome on measure X over the students who teacher j teaches in year $t-1$, our statistical model of student outcomes (e.g. Equation (37)) implies that:

$$y_{i,t-1} - \overline{y}_{j,t-1} = \beta \cdot (X_{i,t-1} - \overline{X}_{j,t-1}) - (\epsilon_{i,t-1} - \overline{\epsilon}_{j,t-1}) \tag{35}$$

Next, note that since $X_{i,t-1}$ is uncorrelated with $\epsilon_{i,t-1}$, we also get that $X_{i,t-1} - \overline{X}_{j,t-1}$ is uncorrelated with $\epsilon_{i,t-1} - \overline{\epsilon}_{j,t-1}$. Thus, 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.

Next, we note that if $\hat{\beta} = \beta$ we get that $\theta_{j,t} = \Theta_j + \nu_{j,t} + \epsilon_{i,t}$. From this, we see that:

$$\hat{\Omega} = \frac{1}{J}\sum \Theta_j\Theta_j' + \Theta_j \cdot (\nu_{j,t-1} + \epsilon_{i,t-1})' + (\nu_{j,t} + \epsilon_{i,t}) \cdot \Theta_j' \tag{36}$$

Under the assumptions regarding the error term, we get that $\frac{1}{J}\sum \Theta_j \cdot (\nu_{j,t-1} + \epsilon_{i,t-1})' \to 0$ and $\frac{1}{J}\sum(\nu_{j,t} + \epsilon_{i,t}) \cdot \Theta_j' \to 0$. Thus, $\hat{\Omega} \to \mathbb{E}[\Theta_j\Theta_j'] = \Omega$.

Similarly, we get that $\mathbb{E}[\theta_{j,t}\theta_{j,t}' - \hat{\Omega} - \frac{1}{N_j}\hat{\Sigma}_\epsilon] = \Omega + \Sigma_\nu + \frac{1}{N_j}\Sigma_\epsilon - \hat{\Omega} - \frac{1}{N_j}\hat{\Sigma}_\epsilon$. So if $\hat{\Omega} \to \Omega$ and $\hat{\Sigma}_\epsilon \to \Sigma_\epsilon$, we get that $\mathbb{E}[\theta_{j,t}\theta_{j,t}' - \hat{\Omega} - \frac{1}{N_j}\hat{\Sigma}_\epsilon] = \Sigma_\nu$ for all j. From method of moments, we can then get that $\hat{\Sigma}_\nu \to \Sigma_\nu$.

□

Theorem 5. *Let θ_0 to be the observed measures for some teacher and θ_1 to be any potential subset of the unobserved measures for the same teacher. Further, let $d(\Theta_j)$ be any policy that a principal would like to implement if Θ_j were fully observed and define $\mathcal{P}(\theta) = \mathbb{E}_{\Theta_j}[d(\Theta_j)|\theta]$. Then:*

$$\mathcal{P}(\theta_0) = \mathbb{E}_{\theta_1}\Big[\mathcal{P}(\theta_1, \theta_0) \;\Big|\; \theta_0\Big]$$

for any θ_0.

Proof. The proof follows almost directly from the law of iterated expectations. Specifically, note that:

$$\begin{aligned}
\mathcal{P}(\theta_0) &= \mathbb{E}_{\Theta}\Big[d(\Theta) \mid \theta_0\Big] \\
&= \mathbb{E}_{\theta_1}\Big[\mathbb{E}_{\Theta}\big[d(\Theta) \mid \theta_1, \theta_0\big] \;\Big|\; \theta_0\Big] \\
&= \mathbb{E}_{\theta_1}\Big[\mathcal{P}(\theta_1, \theta_0) \;\Big|\; \theta_0\Big]
\end{aligned}$$

The first and third equality signs stem from the definition of $\mathcal{P}(\theta_0)$ and $\mathcal{P}(\theta_1, \theta_0)$, while the second uses the law of iterated expectations.

□

C Framework with Teacher Value-Added Drift

C.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:

$$y_{i,t-1} = \beta X_{i,t-1} + \Theta_{j,t-1} + \nu_{j,t-1} + \epsilon_{i,t-1} \tag{37}$$

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 (3) and continuing to denote these estimates $\theta_{j,t-1}$, from this statistical model we get that if $\hat{\beta} \to \beta$:

$$\theta_{j,t-1}|\Theta_{j,t-1} \sim N\Big(\Theta_{j,t-1}, \Sigma_\nu + \frac{1}{N_j}\Sigma_\epsilon\Big) \tag{38}$$

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

$$\Theta_{j,t-1}|\theta_{j,t-1} \sim N\Big(\Omega_j^* \theta_{j,t-1}, \Sigma_j^*\Big) \tag{39}$$

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 III.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:

$$\begin{bmatrix} \Theta_{j,t} \\ \Theta_{j,t-1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega & \mathbb{C}_1 \\ \mathbb{C}_1 & \Omega \end{bmatrix} \right) \tag{40}$$

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:

$$\Theta_{j,t}|\Theta_{j,t-1} \sim N\Big(\mathbb{C}_1\Omega^{-1}\Theta_{j,t-1}, \Omega - \mathbb{C}_1\Omega^{-1}\mathbb{C}_1\Big) \tag{41}$$

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 (41) and (39) to get that:[38]

$$\Theta_{j,t}|\theta_{j,t-1} \sim N\Big(\mathbb{C}_1\Omega^{-1}\Omega_j^*\theta_{j,t-1}, \Omega - \mathbb{C}_1\Omega^{-1}\mathbb{C}_1 + \mathbb{C}_1\Omega^{-1}\Sigma_j^*\Big) \tag{42}$$

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:

$$\hat{\Theta}_{j,t} = \mathbb{C}_1(\Omega + \Sigma_j)^{-1}\theta_{j,t-1} \tag{43}$$

C.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 $\mathbb{C}_1 = \mathbb{V}\big(\Theta_j\big) + \rho\mathbb{V}\big(\phi_{j,t}\big)$

[38] 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 $\mathbb{C}_1\Omega^{-1}\eta + \epsilon$.

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.1 and A.2, 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.1 we illustrate that the components explain a similar percentage of the overall variance regardless of the lag we use. In Figure A.2, 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.

D 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.

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.,:

$$\gamma = \lim_{N\to\infty} (\Theta'\Theta)^{-1}\Theta'\tilde{\Theta} \tag{44}$$

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

$$\hat{\gamma} = \lim_{N\to\infty} (\hat{\Theta}'\hat{\Theta})^{-1}\hat{\Theta}'\tilde{\Theta} \tag{45}$$

where the j^{th} row of $\hat{\Theta}$ is $\hat{\Theta}_j'$ and $\hat{\Theta}_j = \Omega_j^*\theta_{j,t-1}$. Our question is how $\hat{\gamma}$ and γ are related and, more specifically, under what assumptions are they equal.

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 III.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:

$$\hat{\gamma} = \mathbb{E}[\Omega_j^*(\Omega + \Sigma_j)\Omega_j^{*'}]^{-1}\mathbb{E}[\Omega_j^*\Omega]\gamma \tag{46}$$

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:

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

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. $\hat{\gamma} \neq \gamma$.[39] 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.[40]

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

[39] 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.

[40] 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.

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.[41]

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 essence, our comments above are simply noting that the estimates of $\mathbb{E}[\Theta_j|\theta_{j,t}]$ used in the regression should include: a) only the subset of $\theta_{j,t-1}$ measures used in the regression; b) all of the $\theta_{j,t-1}$ measures used in the regression; and c) any additional covariates used in the regression.

E Implied Weights on the Raw Effect Estimates

Note that there were two sets of weights that we discussed in Section IV. 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^*.[42] 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 III. 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.

[41] 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.

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

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 6 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.[43] 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 6), 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.

For middle school, (panel (B) of Table 6), 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.[44] 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.

[43] 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.

[44] 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.

F Incorporating Multiple Years of Data into the Estimates

F.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:

$$\overline{\theta}_{j,-t} = \sum_{m=1}^{M} \theta_{j,t-m} \tag{47}$$

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

$$\mathbb{E}\left[\Theta_{j,t}|\overline{\theta}_{j,-t}\right] = \Omega^{*}\overline{\theta}_{j,-t} \tag{48}$$

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

$$\Sigma_j = \frac{1}{M}\Sigma_\nu + \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N_{j,t-m}}\Sigma_\epsilon \tag{49}$$

where $N_{j,t-m}$ is the number of students teacher j taught in year $t-m$.[45]

As we discuss in Appendix C, 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

[45] 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 F. 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.

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 F 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.

F.2 With Drift

We next use the model presented in Appendix C 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:

$$\begin{pmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{pmatrix}\Bigg|\begin{pmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix}\sim N\left(\begin{bmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{bmatrix},\begin{bmatrix}\Sigma_{j,t-1} & 0\\ 0 & \Sigma_{j,t-2}\end{bmatrix}\right) \tag{50}$$

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.

Next, from our assumptions on drift, we get that

$$\begin{pmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix}\sim N\left(\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix}\right) \tag{51}$$

Thus, from Bayes' Law we get that:

$$\mathbb{E}\left[\begin{pmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix}\Bigg|\begin{pmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{pmatrix}\right]=\begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix}\begin{bmatrix}\Omega+\Sigma_{j,t-1} & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega+\Sigma_{j,t-2}\end{bmatrix}^{-1}\begin{bmatrix}\theta_{j,t-1}\\ \theta_{j,t-2}\end{bmatrix} \tag{52}$$

Finally, from our assumptions on drift we get that:

$$\begin{pmatrix}\Theta_{j,t}\\ \Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix}\sim N\left(\begin{bmatrix}0\\ 0\\ 0\end{bmatrix},\begin{bmatrix}\Omega & \mathbb{C}_1 & \mathbb{C}_2\\ \mathbb{C}_1 & \Omega & \mathbb{C}_1\\ \mathbb{C}_2 & \mathbb{C}_1 & \Omega\end{bmatrix}\right) \tag{53}$$

and so

$$\Theta_{j,t}\Bigg|\begin{pmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{pmatrix}\sim N\left(\begin{bmatrix}\mathbb{C}_1 & \mathbb{C}_2\end{bmatrix}\begin{bmatrix}\Omega & \mathbb{C}_1\\ \mathbb{C}_1 & \Omega\end{bmatrix}^{-1}\begin{bmatrix}\Theta_{j,t-1}\\ \Theta_{j,t-2}\end{bmatrix},\mathbf{\Sigma}\right) \tag{54}$$

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

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

G Teachers with Missing Outcomes

In Section VI, we discussed how to construct empirical Bayes estimates when some of the effect estimates are missing. We now discuss this in more detail.

G.1 Derivation of Empirical Bayes Posterior Distribution

In this section, we provide the derivation of the empirical Bayes' posterior distribution when not all measures have observable estimates, i.e., we derive Equation 16.

To do so, we start by noting that since $\Theta_j \sim N\big(0, \Omega\big)$, we get that:

$$\Theta_{1,j}|\Theta_{2,j} \sim N\Bigg(\Omega_{1,2}\Omega_{2,2}^{-1}\Theta_{2,j}, \Omega_{1,1} - \Omega_{1,2}\Omega_{2,2}^{-1}\Omega_{2,1}\Bigg) \tag{55}$$

We can therefore write that:

$$\Theta_j = \begin{bmatrix}\Omega_{1,2}\Omega_{2,2}^{-1} \\ I\end{bmatrix}\Theta_{2,j} + \eta \qquad \text{with} \qquad \eta \sim N\Bigg(0, \begin{bmatrix}\Omega_{1,1} - \Omega_{1,2}\Omega_{2,2}^{-1}\Omega_{2,1} & 0 \\ 0 & 0\end{bmatrix}\Bigg) \tag{56}$$

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 III to get that:

$$\Theta_{2,j}|\theta_{2,j,t-1} \sim N\Bigg(\Omega_{2,2}\big(\Omega_{2,2} + \Sigma_{j,2,2}\big)^{-1}\theta_{2,j,t-1}, \big(\Omega_{2,2}^{-1} + \Sigma_{j,2,2}^{-1}\big)^{-1}\Bigg) \tag{57}$$

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:

$$\Theta_{2,j} = \Omega_{2,2}\big(\Omega_{2,2} + \Sigma_{j,2,2}\big)^{-1}\theta_{2,j,t-1} + \zeta \qquad \text{with} \qquad \zeta \sim N\Bigg(0, \big(\Omega_{2,2}^{-1} + \Sigma_{j,2,2}^{-1}\big)^{-1}\Bigg) \tag{58}$$

We can then plug in Equation (58) into (56) to get that:

$$\Theta_j = \begin{bmatrix} \Omega_{1,2}\Omega_{2,2}^{-1} \\ I \end{bmatrix} \Omega_{2,2}\big(\Omega_{2,2} + \Sigma_{j,2,2}\big)^{-1}\theta_{2,j,t-1} + \begin{bmatrix} \Omega_{1,2}\Omega_{2,2}^{-1} \\ I \end{bmatrix} \zeta + \eta \tag{59}$$

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 (59) as:

$$\Theta_j|\theta_{2,j,t-1} = N\left(\begin{bmatrix} \Omega_{1,2} \\ \Omega_{2,2} \end{bmatrix} \big(\Omega_{2,2} + \Sigma_{j,2,2}\big)^{-1}\theta_{2,j,t-1}, \begin{bmatrix} \Omega_{1,2}\Omega_{2,2}^{-1} \\ I \end{bmatrix} Var(\zeta) \begin{bmatrix} \Omega_{1,2}\Omega_{2,2}^{-1} \\ I \end{bmatrix}' + Var(\eta) \right) \tag{60}$$

and Equations (58) and (56) make clear that $Var(\zeta) = \big(\Omega_{2,2}^{-1} + \Sigma_{j,2,2}^{-1}\big)^{-1}$ and $Var(\eta) = \begin{bmatrix} \Omega_{1,1} - \Omega_{1,2}\Omega_{2,2}^{-1}\Omega_{2,1} & 0 \\ 0 & 0 \end{bmatrix}$.

G.2 Comparison with Imputation Approach

As mentioned in the Section VI, the most natural alternative approach is to impute the missing values and then construct the empirical Bayes estimates according to Section III. Here we contrast the empirical Bayes approach outlined in Section VI 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 $\mathbb{E}[\theta_{i,t}^k]$, if $\theta_{i,t}^k$ is the value that is missing. Note that the measures are normalized so that $\mathbb{E}[\theta_{i,t}^k] = 0$ for all k. Of course, this approach is problematic as it is does not distinguish between $\theta_{i,t}^k$ 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 $\theta_{i,t}^k$, 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 $\mathbb{E}[\theta_{i,t}^k|\theta_{i,t}^{-k}]$, where $\theta_{i,t}^{-k}$ is the set of measures which are not missing, before calculating the empirical Bayes estimates according to Section III. Note that $\mathbb{E}[\theta_{i,t}^k|\theta_{i,t}^{-k}]$ are themselves the empirical Bayes estimates, so among other things this approach is more complex to implement than the approach mentioned in Section VI. It also means that, in some sense, the approach shrinks the estimates twice: first when constructing $\mathbb{E}[\theta_{i,t}^k|\theta_{i,t}^{-k}]$ 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 VI; 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 III. Figure A.3 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.