The Importance of Student-Teacher Matching: A Multidimensional Value-Added Approach*

Tom Ahn†
Naval Postgraduate School

Esteban M. Aucejo‡
Arizona State University

Jonathan James§
California Polytechnic State University

August 27, 2025

Abstract

We propose a framework for value-added models that flexibly characterizes heterogeneous teacher productivity based on multidimensional student characteristics. We show that teacher effectiveness heavily depends on the specific attributes of their students. For example, the difference in value-added between well-matched and poorly-matched students for the average teacher is approximately 0.1 standard deviations in test scores. Notably, these matching effects are particularly pronounced among low-achieving students. In language arts, the standard deviation in teacher value-added is one-third larger for low-achieving students compared to high-achieving students.

Keywords: value-added, teacher, productivity, matching, multivariate shrinkage

JEL Classification Codes: I21, I24, J21

*Previously titled: The Importance of Matching Effects for Labor Productivity: Evidence from Teacher-Student Interactions. We thank seminar participants at the CEP-LSE education group, University of Pittsburgh, University of Michigan, Michigan State University, Washington University at St. Louis, Association for Education Finance and Policy Conference, and 15th Annual All-California Labor Economics Conference.

†Department of Defense Management, Naval Postgraduate School. Email: sahn1@nps.edu

‡Department of Economics, W.P. Carey School of Business, Arizona State University & CEP & NBER. Email: Esteban.Aucejo@asu.edu

§Department of Economics, California Polytechnic State University. Email: jjames04@calpoly.edu

1 Introduction

Over the last two decades, researchers have coalesced around value-added models (VAM) as the leading framework for evaluating a teacher's impact on students' standardized test score gains. VAMs have been used to inform the public about teacher quality and to advocate for personnel actions. However, these models generally uncover a measure of teacher effectiveness that is assumed to be constant across students. While this framework may be reasonable for studying a teacher's contribution for their representative student, it abstracts away from potential student-teacher matching effects, limiting deeper insights into teacher productivity.[1] Expanding VAMs to consider these matching effects is essential for understanding how teacher effectiveness varies across different students. This could lead to improved student outcomes through better student-teacher assignments and more accurate teacher rankings that reflect differential effectiveness across student groups.

Student-teacher matching has been studied extensively outside the VAM context. For example, Lusher et al. (2018); Gershenson et al. (2018); Gong et al. (2018); Lavy (2015) document that certain teachers are more effective when teaching students of the same gender or race, while Steinberg and Garrett (2016); Aucejo et al. (2019); Graham et al. (2020) show that some teachers have a pedagogical style better suited for high-achieving classes. However, each of these studies examines matching in only one (or a few) dimension(s) and focuses on estimating partial match effects based on observable teacher characteristics, which reflect the average in the population, rather than identifying the match effects of individual teachers. Contemporaneous studies, such as Bates et al. (2022); Delgado (2021); Biasi et al. (2021), have started incorporating match effects into the VAM framework. However, like the earlier literature, these studies have focused primarily on matching in a single dimension at a time.

In this paper, we propose a new estimation framework for multidimensional VAMs. Our

[1] Exceptions include Jackson (2013), which separates value-added (VA) into a teacher and a teacher-school match components, and Abdulkadiroğlu et al. (2020), which allow for match effects between students and schools, where some schools are more effective for specific types of students. Gilraine and Pope (2020) decompose VA into short-run and long-run components to capture teacher contributions to learning that persists over time, while Petek and Pope (2021) shows teachers impacting non-test score student ability.

framework integrates several elements from both prior and contemporaneously developed VAMs but extends them in important ways. First, it is exceptionally well-suited to incorporate multidimensional student-teacher matching. Second, our maximum likelihood-based estimator allows us to relax the commonly imposed distributional assumptions on the value-added distribution. Third, while previous work has primarily focused on the posterior means of value-added estimates (best linear unbiased predictors), our estimator recovers the full posterior distribution, enabling a more thorough analysis of estimate precision for policy analysis.

Because our framework significantly broadens the range of value-added models that researchers can estimate, we provide an extensive discussion on model selection (Section 4.2). Model selection can be oriented either toward the population distribution of value-added components, which is useful for questions related to understanding key aspects of teacher productivity, or based on predictive fit, such as improving value-added estimates for conditional assignments of particular teachers to students. We provide guidance on how to select the most appropriate model for either goal.

We estimate our model using reading and math scores from fourth- and fifth-grade classrooms in North Carolina, incorporating matching across all available observable student characteristics. Our findings indicate that match effects are a significant component of teacher productivity, accounting for 12% of value-added in math and 25% of value-added in reading. For the average teacher, the difference in test scores between a well-matched and a poorly-matched student is at least 0.1σ test score units for both reading and math. We also find that students' prior test scores are the primary driver of matching, accounting for approximately 75% of match effects. Notably, match effects are especially important for low-achieving students. For instance, the standard deviation in teacher value-added is one-third higher for students with prior scores 1 standard deviation (sd) below the mean compared to those 1 sd above the mean. These results highlight greater variability in teacher effectiveness for underperforming students, suggesting that schools with more low-achieving students

should be particularly careful in avoiding poorly matched teachers who could further disadvantage these students. However, once effective teachers are identified, they can produce substantial test score gains for these students.

Using our estimates, we conduct counterfactual simulations to measure potential gains from reallocating teachers to classrooms in a way that maximizes matching effects. On average, students would gain about 0.05 sd in test score units if teachers were optimally assigned within the same district. The benefits are even greater for traditionally disadvantaged populations, with Black male students, for example, achieving average gains of around 0.07 sd in math and reading. The variation in reallocation gains depends on the heterogeneity of student composition, differences in match coefficients among teachers, and the extent to which teachers are already optimally allocated.

Finally, we explore problems with ranking teachers based on traditional value-added models. We demonstrate that a teacher's effectiveness is significantly influenced by the students they are assigned, which is not considered when constructing rankings based on VAMs in the past. Our results show that 29% (60%) of the teachers in the bottom 5% for reading value-added would no longer rank in that group if they were assigned to a better-matched classroom within their school (district).[2]

2 Data

We use administrative records maintained by the North Carolina Education Research Data Center. This longitudinal database provides information on teachers and students from all public schools in the state. Teachers are matched to students, and students are followed year-to-year and grouped into classes, as long as they attend a NC public school.

Our sample covers grades 4 and 5 in the academic years 2007/08 to 2013/14. We focus on teachers with "self-contained" classrooms, defined as a group of students receiving math

[2]Hanushek (2009) and Chetty et al. (2014) explore ranking to replace less effective teachers. Condie et al. (2014) analyzes how teacher heterogeneity biases value-added, leading to incorrect rankings.

Table 1: Summary Statistics

	Mean	Std. Dev.		Mean	Std. Dev.
Panel A: Student-Year (N=804,879)			Panel B: Classroom-Year (N=44, 738)		
Female	0.499	0.500	Female Teacher	0.904	0.295
Black	0.258	0.438	Black Teacher	0.125	0.331
FRL-status	0.523	0.499	Teacher Experience	11.065	8.643
LEP	0.057	0.233	Class Size	17.9	4.92
Lagged Reading Score	0.045	0.972			
Lagged Math Score	0.054	0.972			
Panel C: School-Year (N=7,748)			Panel D: District-Year (N=751)		
Enrollment	103.9	64.3	Enrollment	1071.7	1958.7
Teacher count	5.8	3.2	Teacher count	59.7	107.5

Note: Summary statistics correspond to grades 4 and 5 in 2007/08 to 2013/14 for the analytic sample. Approximately 50 percent of the raw data is retained, after deleting observations due to missing student characteristics, non-self-containing classes, and other irregularities. Appendix A provides a detailed explanation of how we constructed our analytic sample. Panels A to D show student-year, classroom-year, school-year, and district-year summary statistics, in order. Teacher experience refers to the period when they draw a salary from the NC Department of Public Instruction. FRL is free or reduced-price lunch which is a proxy for economic disadvantage, LEP is limited English proficiency. Reading and math scores have been standardized with mean zero and standard deviation one by grade and year. Class % denotes classroom characteristics for a teacher-year observation.

and reading instruction from the same teacher and only that teacher in both subjects. We exclude teachers assigned to more than one set of reading and math classes.

We define our sample in this restrictive way to minimize spill-overs. Teachers assigned to more than one self-contained classroom may split attention and effort. If students are taught by more than one teacher, it becomes difficult to apportion credit for test score gains. Defining the self-contained classroom in this way also precisely and narrowly defines a student's peers, ensuring interactions among teacher, student, and classmates are not polluted by other agents, allowing us to cleanly isolate matching effects.

Panel A presents annual student information. About half of students are females, 26% are Black, 52% receive free or reduced-price lunch (FRL), and 6% have limited English proficiency (LEP). Panel B reports class-level statistics, where the average class seats about 18 students. About 90% of teachers in those classrooms are females, 13% are Black, and on average, they have over 11 years of teaching experience. Panels C and D show the spread of teachers and students within the approximately 1,100 schools over 110 school districts. The average school has 6 teachers with about 100 4th and 5th-grade students. The average district contains about 10 schools, 60 teachers, and 1,100 students. Districts vary greatly

in size. The smallest district has a single school with less than 20 students, taught by one teacher. The largest district contains 106 schools with approximately 850 teachers and 16,000 students.

3 Conceptual Framework and Methods

3.1 Multivariate Value-added Model

We are interested in modeling the determinants of the test score of student i in year t when assigned to teacher j, A_{it}, which we assume can be represented as:

$$A_{it} = x'_{it}\beta + \alpha_{ijt} + \nu_{it} \tag{1}$$

$$\text{where} \quad \alpha_{ijt} = z_{it1}\delta_{j1} + z_{it2}\delta_{j2} + \cdots + z_{itK}\delta_{jK}$$

The first term in Eq. (1) captures aspects of student test scores that are deterministic and not dependent on a specific assignment of a student to a teacher. The vector x_{it} represents observed student, teacher, classroom, school, and district attributes, while β is the influence of these attributes on test scores. The second term, α_{ijt}, represents teacher j's contribution to test scores that is specific to student i in year t; broadly speaking, it represents the teacher's value-added. The final component, ν_{it}, is an idiosyncratic shock to student test scores that is independent of the other variables in the model.

Our model aims to characterize a teacher's value-added in a flexible way, which we define as a linear function of teacher-specific coefficients and student characteristics, as reflected in Eq. (1). Teacher productivity depends on multiple student characteristics, $\{z_{itk}\}_{k=1}^{K}$, which may include prior academic achievement and demographics, as well as year indicators to capture time-varying teacher-level effects, including non-linearities and interactions among these variables. The teacher matching coefficients, $\{\delta_{jk}\}$, reflect how the teacher's value-added changes given a change in the student attribute, which equates to a measure of teacher

j's comparative advantage in teaching students with attribute k. The variance of δ_{jk} describes the degree of heterogeneity in the teacher population for teaching students with attribute k. Therefore, uncovering this variance will be the primary basis for determining how important student-teacher assignments are to overall student test scores.

Equation (1) nests several models in the literature. For example, Koedel et al. (2015) discusses a univariate VAM in which a teacher's contribution to test scores is constant over time and equal for all students (i.e., $\alpha_{ijt} = \alpha_j$ where $K = 1$ and z_{it1} is a constant). Chetty et al. (2014) lets the teacher's contribution to test scores to change over time but assumes it is the same for all students within a given year (i.e., $\alpha_{ijt} = \alpha_{jt}$, with K as the number of years in the data and z_{itk} as an indicator variable if i is assigned to j in year k).

Our framework encompasses models from Bates et al. (2022); Delgado (2021); Biasi et al. (2021), where teacher effectiveness is separately estimated for students split into mutually exlclusive subpopulations. For example, to have value-added differ by student FRL status, we can set $K = 2$ and define z_{it1} and z_{it2} as indicators that i is FRL or non-FRL eligible, respectively, making δ_{j1} and δ_{j2} teacher j's value-added for FRL and non-FRL students. Our model also has unique advantages for multidimensional matching. First, Eq. (1) accommodates both continuous and discrete matching variables. In a subpopulation approach, continuous variables must be discretized. Second, with high dimensional matching, for example with K student dimensions, splitting the data into subpopulations entails at a minimum 2^K teacher parameters, while Eq. (1) could be written with as few as $K + 1$ parameters.

3.2 Estimation

For teacher j, let $\delta_j = [\delta_{j1} \quad \delta_{j2} \quad \delta_{j3} \quad \cdots \quad \delta_{jK}]'$ denote their $K \times 1$ vector of teacher coefficients from Eq. (1). Our goal is to recover the population distribution of these coefficients, where $f(\delta|\Psi)$ is the probability density function, characterized by the unknown parameters Ψ. The properties of this distribution are of central importance and characterize the contribution of each attribute to value-added in the population. In addition, for many research

questions, it is useful to have estimates of the teacher coefficients to compare across teachers. Rather than focus on noisy point estimates of δ_j, the population distributions can be used to form empirical Bayes estimators or best linear unbiased predictors (BLUPs). If A_j denotes all observed test score outcomes for teacher j, then the empirical Bayes' estimator is defined as $\mathrm{E}(\delta_j|A_j, \Psi)$, where the parameters of the population distribution serve as the prior.[3]

A key contribution of our estimator is that we place minimal assumptions on the distribution of the teacher coefficients. A common approach in the literature is to assume that the value-added coefficients are normally distributed (i.e., $\delta \sim N(\gamma, \Delta)$ and $f(\delta|\Psi)$ represents the density function of a normal distribution with $\Psi = \{\gamma, \Delta\}$ containing the mean and variance parameters). Even papers that do not impose normality, for example, Chetty et al. (2014), typically limit their estimators to the class of distributions of empirical Bayes' estimators that are linear functions of the data, a property that is consistent with a normal distribution, but not true of all distributions in general.

The normality assumption has two major drawbacks. First, it limits our ability to understand the true nature of these coefficients. Distinguishing among distributions with skew, many outlying observations, multiple modes, or relatively flat surfaces is directly relevant to understanding the impact teachers have on student test scores. These distinctions become even more important in our setting, where the components of value-added are multidimensional, and the relationship among the variables can potentially take many complex shapes.

Second, it may bias the empirical Bayes estimators discussed above because it imposes the wrong prior. Recently, Gilraine et al. (2020), using a maximum likelihood procedure to estimate a non-parametric univariate value-added model, shows that if the true value-added distribution is non-normal, then imposing normality in the empirical Bayes estimator can lead to inaccurate estimates. This issue may become even more prevalent in our setting where value-added is multidimensional, and potentially many of the components of the underlying value-added process may be poorly approximated by a normal distribution.

[3]It is also common in the literature to construct empirical Bayes' estimators using only a subset of the data, which can be denoted as $\mathrm{E}(\delta_j|\widetilde{A}_j \subset A_j, \Psi)$.

To allow for a more flexible distribution and avoid misspecification, we extend the value-added literature and model the population distribution of δ using a C component mixture of normals, which can approximate a wide range of distributions. Specifically, the probability density function of these coefficients is $f(\delta|\Psi) = \sum_{c=1}^{C} \pi_c \phi(\delta|\gamma_c, \Delta_c)$, where C is the number of components in the mixture, π_c is the share parameter for each component, and $\phi(\delta|\gamma_c, \Delta_c)$ is the probability density function of a multivariate normal with component-specific mean γ_c and covariance Δ_c. The Gaussian mixture relaxes the assumption of normality in a way that can be highly tailored to each research setting. Our approach is particularly useful when there is limited data or when the dimensionality of the model is large, forcing the researcher to be conservative with the number of parameters. For example, the fact that δ is multidimensional in our framework prevents us from implementing the non-parametric approach developed by Gilraine et al. (2020) due to the curse of dimensionality.[4] Using Gaussian mixtures, the researcher can choose a smaller number of components for the mixture, and reduce the number of parameters even further if necessary by placing constraints on the covariance matrices, Δ, for example, by assuming a shared covariance matrix across all components or imposing diagonal covariances. With larger data sets, parsimony is less important and a larger number of components can be used with more flexible covariance matrices.

Replacing the summation in Eq. (1) with a vector product, the estimating equation is:

$$A_{it} = x'_{it}\beta + z'_{it}\delta_j + \nu_{it} \tag{2}$$

where z_{it} is a $K \times 1$ vector of attributes that determine value-added for student i in year t.

While Eq. (2) shares some similarities with a traditional linear mixed model, there are important limitations that make employing off-the-shelf estimators undesirable in this context. First, as discussed in Chetty et al. (2014), a traditional random effects approach may lead to biased estimates of β because of possible non-random assignment of students

[4]Gilraine et al. (2020) uses 5,000 points to approximate the distribution in a level-only VAM. Extending to the multidimensional setting quickly becomes infeasible. In our empirical context, where each teacher is associated with 19 coefficients, we would need to estimate $5,000^{19} = 1.9e70$ parameters.

to teachers. Second, standard mixed models assume normality of the random coefficients, which as discussed may be undesirable in this setting. Finally, the large-scale nature of the problem (i.e., high-dimensional integration on a large dataset) requires a tailored solution to ensure computational speed and numerical stability.

Our estimation strategy is similar to the two-step approach in the value-added literature. The parameters in β are treated as nuisance parameters and estimated using a fixed effect estimator in a first stage. Given unbiased estimates of β, residuals are formed that, by construction, are only a function of the value-added components and the idiosyncratic shock to test scores. These residuals serve as noisy observations of the teacher coefficients and are the basis for the estimation of the population parameters Ψ in a second stage.

To formalize the estimator, let n_j denote the number of students assigned to teacher j across the entire dataset. Define the following design matrices: A_j, a $n_j \times 1$ vector of student test scores for teacher j, X_j, a matrix of observed student, classroom, and district characteristics whose rows contain the x_{it}'s associated with teacher j, Z_j, a $n_j \times K$ matrix containing the determinants of the teacher's value-added where rows contain the z_{it}'s associated with teacher j, and finally ν_j, a $n_j \times 1$ vector of test score residuals.

Given $\hat{\beta}$, the vector of test score residuals for each teacher is defined as:

$$Y_j = A_j - X_j\hat{\beta} = Z_j\delta_j + \nu_j$$

Under the classical assumptions that the error term ν_{it} is independent and normally distributed with $\text{Var}(\nu_{it}) = \sigma^2$, the vector of residuals for each teacher, conditional on δ_j, will follow a multivariate normal distribution:

$$Y_j|\delta_j \sim N(Z_j\delta_j, \sigma^2 I_{n_j}) \tag{3}$$

Where I_{n_j} is the $n_j \times n_j$ identity matrix.

In the second step of the estimator, we use these residuals to estimate the underlying

population distribution $f(\delta|\Psi)$. We depart from the previous literature and use maximum likelihood to estimate the parameters of the Gaussian mixture. Maximum likelihood estimation is an attractive approach to these problems because of its simplicity, efficiency, and flexibility. However, maximum likelihood requires an iterative algorithm to optimize the log-likelihood function, which could be computationally burdensome given the complexity of our multidimensional model and because these models are often estimated on large, administrative datasets. To address this, rather than form the likelihood based on the distribution of the residuals in Eq. (3), we project these residuals onto attributes that determine value-added, Z_j.[5] As we show in Appendix B, projecting the residuals onto the student attributes yields least squares estimates of the teacher coefficients, δ_j^{LS}, which, given the properties of the least squares estimator, are unbiased and distributed $\hat{\delta}_j^{LS} \sim N(\delta_j, \hat{\Sigma}_j)$, where $\hat{\Sigma}_j$ is the variance-covariance matrix of the least squares estimates.[6] The least squares estimates serve as noisy signals of each teacher's true coefficients from which we can use the sampling distribution to form an integrated log-likelihood function to estimate the population parameters of the distribution of the teacher coefficients:

$$\widehat{\Psi} = \underset{\Psi}{\operatorname{argmax}} \sum_{j=1}^{J} \ln \left(\int h(\hat{\delta}_j^{LS}|\delta, \hat{\Sigma}_j) f(\delta|\Psi) d\delta \right) \tag{4}$$

Where $h(\cdot)$ is the density function of the least squares estimate and its relationship to the true underlying teacher coefficients δ_j, and $f(\delta|\Psi) = \sum_{c=1}^{C} \pi_c \phi(\delta|\gamma_c, \Delta_c)$ is the population distribution of the teacher coefficients and $\Psi = \{\pi_c, \gamma_c, \Delta_c\}_{c=1}^{C}$ are its parameters.

The main purpose of forming the estimator on the least squares estimates rather than the residuals is that it reduces the dimension of the teacher data from n_j to K, making

[5] This approach is very similar to other papers in the literature, where most valued-added estimators are based on the average of the teacher residuals, which is equivalent to projecting the residuals on a constant.

[6] If the Z_j in $\delta_j^{LS} = (Z_j'Z_j)^{-1}Z_j'Y_j$ is not full rank, a subset of linearly independent columns of Z_j are used for the projection. In these cases, the least squares estimates will either provide estimates of a subset of the teacher coefficients or estimates of a linear combination of the teacher coefficients, where the mapping is known. The projection is only used to reduce the dimensionality of the data and does not affect identification.

estimation of the model more tractable.[7] We optimize the likelihood in Eq. (4) using a modified expectation-maximization (EM) algorithm. The EM algorithm (Dempster et al., 1977) is a popular method for maximum likelihood estimation of Gaussian mixture models. The main benefits of the EM algorithm are that it is easy to implement, guarantees conformity of parameter constraints, and is globally convergent. The downside of the EM algorithm is that in certain situations, for example, poorly separated mixtures or when the true variance of the match coefficients is zero, the algorithm may have a painfully slow rate of convergence. Unfortunately, practitioners may not know *a priori* how many components to use in the mixture or which characteristics are important for matching and may wish to select their model based on likelihood ratio tests or other information criteria. These test statistics rely on solving Eq. (4) even in these difficult situations. To address the slow rate of convergence of the EM algorithm, we use the modified EM algorithm proposed by Jamshidian and Jennrich (1997) based on Broyden's method, which preserves the desirable features of the EM algorithm but has a significantly faster rate of convergence. We discuss the details of our full two-step estimation approach in Appendix B. We also show how our framework can be adapted to the joint estimation of multiple outcomes in Appendix C.

3.2.1 Posterior Distributions and Empirical Bayes' Estimators

In many settings, in addition to estimating the parameters of the distribution, it is desirable to uncover the likely values of each individual teacher's value-added coefficients.[8] These analyses can be conducted by constructing posterior probability density functions for each teacher. The posterior distributions are based directly on the residuals in Eq.(3) and combines each teacher's observed data with information about the population distribution to create a distribution of plausible values for their value-added coefficients. Because the pos-

[7] While reducing the dimensionality in this way is also possible in the linear mixed model framework, most off-the-shelf implementations do not take this approach, which has significant negative effects on computation time especially when $n_j >> K$ (Bates et al., 2015). The projection significantly improves computation time in our framework since teachers often have many more students than teacher coefficients.

[8] For example, hiring or firing of experienced teachers can be based on predictors of their value-added.

terior distributions potentially rely heavily on the population distribution, if the population distribution is incorrectly specified, then the posterior distributions will be incorrect. Addressing this problem is one of the motivations for using the Gaussian mixture because it provides more flexibility. Given Y_j, Z_j, $\hat{\sigma}^2$, as well as the population parameter estimates, $\widehat{\Psi}$, the posterior distribution for δ_j will also be a C component Gaussian mixture following:[9]

$$
\begin{aligned}
p(\delta_j|Y_j, Z_j, \hat{\sigma}^2, \widehat{\Psi}) &= \textstyle\sum_{c=1}^{C} q_{jc}\phi(\delta|E_{jc}, V_{jc}) \qquad (5)\\
\text{where: } q_{jc} &= \frac{\hat{\pi}_c\phi(Y_j|Z_j\hat{\gamma}_c, Z_j\hat{\Delta}_c Z_j' + \hat{\sigma}^2 I_{n_j})}{\sum_{c'=1}^{C}\hat{\pi}_{c'}\phi(Y_j|Z_j\hat{\gamma}_{c'}, Z_j\hat{\Delta}_{c'} Z_j' + \hat{\sigma}^2 I_{n_j})}\\
E_{jc} &= \hat{\gamma}_c + \hat{\Delta}_c Z_j'(Z_j\hat{\Delta}_c Z_j' + \hat{\sigma}^2 I_{n_j})^{-1}(Y_j - Z_j\hat{\gamma}_c)\\
V_{jc} &= \hat{\Delta}_c - \hat{\Delta}_c Z_j'(Z_j\hat{\Delta}_c Z_j' + \hat{\sigma}^2 I_{n_j})^{-1} Z_j\hat{\Delta}_c
\end{aligned}
$$

The elements describing the distribution in Eq.(5) include, q_{jc}, the probability that teacher j belongs to component c, E_{jc}, the expected value of δ_j given j belongs to c, and V_{jc}, the uncertainty of δ_j given that j belongs to c.

In the extreme case, for a teacher with no data, their posterior distribution will simply be the population distribution. But as data are added their posterior distribution will rely less on the population distribution and in the limit will approach a degenerate distribution with a single mass point at the true value of their coefficients. In a univariate setting, the precision of the posterior is proportional to the number of observations per teacher. However in a multivariate setting, in addition to sample size, the precision of the posterior also depends on the covariability / collinearity of the match characteristics of the students assigned to the teacher, Z, and the covariance of the teacher coefficients, Δ. For example, if the covariance of the match coefficients in the prior, Δ, suggests that teacher value-added

[9]Formally, for any general distribution $f(\delta|\widehat{\Psi})$, these densities take the form:

$$p(\delta_j|Y_j, Z_j, \hat{\sigma}^2, \widehat{\Psi}) = \frac{p(Y_j|\delta, Z_j, \hat{\sigma}^2)f(\delta|\widehat{\Psi})}{\int p(Y_j|\delta', Z_j, \hat{\sigma}^2)f(\delta'|\widehat{\Psi})d\delta'}$$

Where $p(Y_j|\delta, \cdot)$ is the joint density of Y_j given δ.

is highly correlated between two types of students, then because of this correlation, if the teacher is only observed teaching one type of student, these data will also be informative about the teacher's value-added even for the student group that is not observed.[10]

How the distributions in Eq.(5) are used will depend on the purpose of the exercise. For our analysis on the gains to re-allocating teachers across classrooms, we use the entire density to characterize the gains. In other situations, it is common to focus on the expected value of the posterior distribution $\mathrm{E}(\delta_j|Y_j, Z_j) = \sum_{c=1}^{C} q_{jc}E_{jc}$. These expected values are the best linear unbiased predictors of the individual teacher's value-added coefficients.[11] Less common in the literature is to focus on the variance, or precision, of the posterior distribution, $\mathrm{Var}(\delta_j|Y_j, Z_j)$. Incorporating the precision into the analysis would be particularly useful, for example, when ranking teachers based on best linear unbiased predictors (Hanushek, 2009). In the univariate setting, as mentioned, sample size thresholds would be sufficient to ensure a given degree of precision. However, in the multivariate setting, relying on sample size alone would not be enough. In these cases, researchers could make precision thresholds directly from the variance of the posterior distribution to be certain that a sufficient amount of information is used when evaluating teachers based on value-added estimates.[12]

4 Empirical Specification

We motivate our empirical specification by discussing our first stage, model selection, and identification. We also provide preliminary evidence on the importance of match effects.

[10]This feature resembles the approach implemented in Chetty et al. (2014) where predictions of teacher value-added at time t are constructed using data from all periods except t.

[11]Technically because the mean of the posterior distribution is not a linear function of the data, these are the best-unbiased predictors (BUP), which in our case differs from the best linear unbiased predictor. Under the assumption of normality, the BLUP is the BUP.

[12]In practice, most policy proposals employing VAMs center around the tails of the distribution. For teachers with limited information, their values will be more concentrated toward the mean, most likely excluding them from policy action.

4.1 First Stage

The teacher match effects are estimated from test score residuals, where the vector of characteristics, x, are partialled out from the test scores. In our model, x includes all observed student characteristics available in our data: the prior year test scores, gender, race, FRL and LEP status, and interactions among these variables. The model also accounts for yearly averages at the teacher, school, and district levels for each characteristic, as well as both linear and quadratic terms for the annual number of students assigned to the teacher, school, and district. Finally, we control for teacher experience and fixed effects for year and grade.

The estimation of match effects is based on the residuals of test scores after accounting for the effects of x. In the next section, we systematically examine which dimensions of student-teacher matching are most evident in the data. For now, we take z, the vector of teacher coefficients relevant for value-added, as given. To produce an unbiased estimate of β that accounts for sorting of teachers base on their comparative advantage, in our first stage we estimate Eq. (2) with an interacted fixed effect model where the teacher fixed effects are interacted with the matching variables in z.[13]

4.2 Model Selection of Match Effects

To select variables to include in the vector of attributes that determine value-added, z, specifically those related to student-teacher matching, we can draw from established results in the education literature as well as from standard approaches to model selection as guides. As emphasized by Vaida and Blanchard (2005), in the case of clustered analysis, the optimal criterion for selecting a model hinges on whether the main focus is on questions about the population distribution or on the behavior of individual clusters. This dichotomy is at the

[13]In our empirical specification, z contains the matching variables as well as teacher-year effects. In the first stage we estimate β only interacting the teacher fixed effects with the matching variables in z, which does not include the teacher-year effects. We did not include teacher-year fixed effects in the first stage because they would absorb any constant-within-year teacher variables in x, such as teacher experience, or control variables at the classroom, school, or district level. Our goal is to identify teacher value-added while holding these variables fixed, keeping them in x, so we exclude teacher-year fixed effects in the first stage.

forefront of value-added models, where on one hand, the aim may be to provide a deeper understanding of the determinants of student outcomes, while on the other hand, a different analysis may center on forecasting a specific teacher's value-added.

Model selection involves choosing an empirical model that avoids underfitting or overfitting the data. If too few match variables are included, this can lead to biased and misinterpreted estimates. For example, if matching is primarily on lag score, but lag score is omitted, then it will appear that any included variables that is correlated with lag score are more important for matching than they actually are. Furthermore, including too few variables will lead to understating the overall importance of matching by leaving out important features. On the other hand, including too many variables risks overfitting the data, where spurious correlations can be mistaken for genuine match effects. In addition, more variables increases computational burden and may present challenges for interpreting the results, especially if variables enter the model in a complex way.

Additionally, both underfitting and overfitting may lead to poor out-of-sample predictions. Shrinkage estimators have played a crucial role in value-added models to address the issue of overfitting in predicting a teacher's contribution to test scores. As more variables are included into the model, this can lead to less precise estimates of each variable's effect. If this imprecision is too large this will result in an increase in the shrinkage of the estimates, reducing the overall predictive accuracy.

Our main focus is quantifying the importance of match effects in value-added, which pertains to the characteristics of the population distribution of δ. To systematically assess which dimensions of comparative advantage are relevant for our data, we first estimate a traditional VAM that differentiates teachers based solely on their absolute advantage, without match effects. Specifically, $K = 1$ and z includes only a constant intercept. We then sequentially add match components and assess how these changes impact the model's performance across model selection criteria related to population fit as well as predictive fit. We consider possible match effects in all observed student characteristics in our data: lagged

test scores, gender, race, FRL- and LEP-status, and their interactions.

To conduct this analysis, we need a preliminary estimate of β that remains constant across specifications, ensuring that a consistent set of test score residuals, Y_j, are used for the dependent variable to compare models. For this, we use a simple fixed effects model: $A_{it} = x'_{it}\beta + \alpha_j + \varepsilon_{it}$, where α_j is the fixed effect for teacher j, to obtain a preliminary estimate of β. The simple fixed effects model is conventionally appealing because it is the one used in the level-only framework of Chetty et al. (2014). Thus, any improvements in fit from including match variables can be directly compared to the level-only model as a baseline. Additionally, the simple fixed effect model can reasonably account for some, although not all, forms of non-random assignment of students to teachers.[14] Once we have established our main specification for the teacher coefficients, for our main results we re-estimate the first stage following Section 4.1.[15]

We consider three model selection criteria to assess population fit. First, we follow Vaida and Blanchard (2005) and calculate the Akaike Information Criterion (AIC), based on the marginal likelihood:

$$\text{AIC} = -2\sum_{j=1}^{J} \ln\left(\int h(Y_j|Z_j,\delta) f(\delta|\widehat{\Psi}) d\delta\right) + (K^2 + 3K)$$

where $K^2 + 3K$ is the bias correction based on δ following a $K \times 1$ multivariate normal distribution. Vaida and Blanchard (2005) argue that the marginal AIC is the appropriate model selection criterion in mixed effect models for population parameters.

Second, the adjusted R-squared using the least squares estimates $\hat{\delta}_j^{LS}$ can also be used to make comparisons across different models. Defining $\overline{y}$ as the sample average of the score

[14] Teacher fixed effects capture a teacher's average contribution across their assigned students. Therefore, if teachers are assigned based on comparative advantage, the simple fixed effects estimator estimates β using within-teacher variation while holding the teacher's average comparative advantage for their assigned students constant.

[15] In our data, the residuals from the first stage using a simple teacher fixed effect were very similar to those from a model with fully interacted fixed effects and match variables, as described in Section 4.1. In fact, the correlation between the residuals from both models was 0.999 for both math and reading, suggesting minimal differences between the two first-stage approaches.

residuals from the first stage, and $n = \sum_{j=1}^{J} n_j$ as the total number of data points, the adjusted R-squared takes the form:

$$\text{adj-}R^2 = 1 - \left(\frac{\sum_{j=1}^{J} (Y_j - Z_j \hat{\delta}_j^{LS})'(Y_j - Z_j \hat{\delta}_j^{LS})/(n - J \times K)}{\sum_{j=1}^{J} (Y_j - \overline{y})'(Y_j - \overline{y})/(n-1)} \right)$$

While $\hat{\delta}_j^{LS}$ are extremely noisy and would lead to imprecise out-of-sample predictions of the teacher coefficients, they are still unbiased estimates and can be appropriately used to calculate an unbiased estimate of R-squared once we adjust the degrees of freedom for the total number of estimated parameters $J \times K$.

Finally, we directly test each additional match variable included in the model with a likelihood ratio test described by Stram and Lee (1994) on the null hypothesis that the variance of the additional match effect is zero, i.e., $H_0 : \text{Var}(\delta_K) = 0$.

While AIC, adjusted R-squared, and the zero-variance hypothesis test will provide a clear picture of which match effects are most relevant at the population level, we also present additional criteria that center more on each model's predictive performance. For instance, if each teacher were assigned one additional student, how well would each model perform in predicting the test scores for this new data?

We assess each model's ability to forecast out-of-sample using approximate leave-one-out predictive cross-validation following Braun et al. (2012); Marshall and Spiegelhalter (2003). In this approach, we calculate the predictive distribution of the test score residual, y_i, for each student i, where y_i is excluded when forming the predictive distribution. Then, we evaluate the model's ability to forecast y_i using only the remaining data. The predictive distribution is constructed from the posterior distribution of the teacher coefficients in Eq. (5), excluding the data from student i. Specifically, letting y_i denote the test score for individual i, then the predictive distribution for y_i given the model and the remaining data is normally distributed with mean $\mu_i^{(-i)} = z_i' \, \text{E}(\delta_j | Y_j^{(-i)}, \widehat{\Psi})$ and variance $v_i^{(-i)} = z_i' \, \text{Var}(\delta_j | Y_j^{(-i)}, \widehat{\Psi}) z_i + \hat{\sigma}^2$.[16]

[16]This method is an approximate cross-validation because y_i is only left out in forming the predictive distribution, not in re-estimating the model parameters, Ψ. In contrast to a conventional leave-one-out

There are several methods available to evaluate the observed data against the predictive distribution. Here, we focus on assessing predictive accuracy using the mean squared error (MSE), which uses only the mean of the predictive distribution as a point estimate for the outcome. Denoting y_i^{obs} as the observed test score residual, the mean squared error is defined as:[17]

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{obs} - \mu_i^{(-i)}\right)^2$$

Panel A in Table 2 report the four model selection criteria discussed above for 17 different models fit separately for student math and reading scores, where each row corresponds to adding one additional match variable to the prior fit. There are three main takeaways from these results. First, match effects are clearly important. For math, the inclusion of match effects increases the adjusted R-squared by 2 percentage points, which is about half as big as the increase in adjusted R-squared by adding year effects. For reading, R-squared increases by 1.5 percentage points, which is larger than the increase in adjusted R-squared from the addition of year effects. Second, student-teacher matching is occurring across every observed student characteristics. For models $K = 1$ to 7, where the characteristics are additively separable, there is a consistent improvement in adjusted R-squared, AIC, and we clearly reject the null hypothesis that the variance of the match coefficients are zero. However, in both subjects, for models $K = 12$ to 17, many of the match effects on the interaction of the student characteristics are not statistically significant. Finally, the results suggest that the

cross-validation where y_i is left out and the model is re-estimated each time, for the approximate approach, the model is only estimated once, which is more feasible for large datasets.

[17] Alternatively, proper scoring rules can be used to compare the observed data to the entire predictive distribution, rather than just its mean. One widely used approach is *log scoring*, which incorporates the precision of the predictive distribution by evaluating the observed data against the logarithm of the predictive density. The log score takes the form:

$$\text{LS} = \frac{1}{n}\sum_{i=1}^{n} -\frac{1}{2}\left(\ln(2\pi) + \ln\left(v_i^{(-i)}\right) + \frac{1}{v_i^{(-i)}}\left(y_i^{obs} - \mu_i^{(-i)}\right)^2\right)$$

where data points predicted with high precision but far from their expected values receive a larger penalty than similarly distant points with less precise predictions. Braun et al. (2012) argue that log scoring is asymptotically equivalent to the conditional AIC, which Vaida and Blanchard (2005) propose for assessing cluster-level performance in mixed models.

preferred model depends on the analysis focus. If the goal is to understand the population fit, models with more match effects are preferable. However, if the analysis is primarily aimed at prediction, a simpler model with fewer match effects might be adequate. For example in math, AIC is maximized at $K = 13$, but there is no improvement in MSE after $K = 7$.

Because we aim to make comparisons between results in math and reading, applying a consistent model to both datasets will make the comparison more accurate. Evaluating the results for the population fit for math in Panel A in Table 2 shows that AIC is maximized for math at $K = 13$, and we fail to reject the null that the variances are zero at the 1% level for models greater than $K = 13$. For reading, AIC is maximized at $K = 11$. Given that the adjusted R-squared for reading is maximized at $K = 13$ and that there are no changes in MSE when increasing from $K = 11$ to 13, we chose to use $K = 13$ for the match effects in our empirical specification for both subjects.

In addition to the match effects, we also include teacher year effects, which corresponds to teacher-year specific intercept terms. Since our data spans years 2008 to 2014 this adds six additional parameters to the vector of teacher coefficients, increasing K to 19. Thus, in our model, the value-added of teacher j for student i in period t is:

$$\begin{aligned} VA_{ijt} =& \delta_{jt} + (A_{i(t-1)})\delta_{j8} + (A^2_{i(t-1)})\delta_{j9} + (\text{FEM}_i)\delta_{j10} + (\text{BL}_i)\delta_{j11} \\ &+ (\text{FRL}_{it})\delta_{j12} + (\text{LEP}_{it})\delta_{j13} + (A_{i(t-1)} \cdot \text{FEM}_i)\delta_{j14} \\ &+ (A_{i(t-1)} \cdot \text{BL}_i)\delta_{j15} + (A_{i(t-1)} \cdot \text{FRL}_{it})\delta_{j16} \\ &+ (A_{i(t-1)} \cdot \text{LEP}_{it})\delta_{j17} + (\text{FEM}_i \cdot \text{BL}_i)\delta_{j18} \\ &+ (\text{FEM}_i \cdot \text{FRL}_{it})\delta_{j19} \end{aligned} \tag{6}$$

Where $A_{i(t-1)}$ is the same subject lagged test score, FEM_i is an indicator variable if the student is female, BL_i is an indicator if the student is Black, FRL_{it} is an indicator for the student's free or reduced-price lunch status in year t, and LEP_{it} indicates if the student is designated as limited English proficiency in year t. The coefficient, δ_{jt} represents a year-

Table 2: Model Selection Criteria

		Math				Reading			
Model	K	AIC	adj-R^2	LR	MSE	AIC	adj-R^2	LR	MSE
Panel A: Sequentially Estimated Matching Models Assuming a Multivariate Normal Distribution									
Intercept Only	1	0	0.1308		0.2541	0	0.0456		0.2852
+ A_{t-1}	2	-1279	0.1390	0.0000	0.2536	-1109	0.0522	0.0000	0.2850
+ A_{t-1}^2	3	-3406	0.1449	0.0000	0.2529	-1704	0.0569	0.0000	0.2849
+ FEM	4	-3462	0.1458	0.0000	0.2529	-1761	0.0570	0.0000	0.2849
+ BL	5	-3544	0.1467	0.0000	0.2529	-1799	0.0578	0.0000	0.2848
+ FRL	6	-3643	0.1477	0.0000	0.2528	-1853	0.0587	0.0000	0.2848
+ LEP	7	-3698	0.1487	0.0000	0.2528	-1869	0.0590	0.0000	0.2848
+ (A_{t-1} × FEM)	8	-3748	0.1493	0.0000	0.2528	-1950	0.0602	0.0000	0.2848
+ (A_{t-1} × BL)	9	-3822	0.1499	0.0000	0.2528	-2056	0.0613	0.0000	0.2849
+ (A_{t-1} × FRL)	10	-3870	0.1503	0.0000	0.2528	-2192	0.0629	0.0000	0.2849
+ (A_{t-1} × LEP)	11	-3894	0.1506	0.0000	0.2528	-2305	0.0634	0.0000	0.2849
+ (FEM × BL)	12	-3926	0.1507	0.0000	0.2528	-2292	0.0635	0.3427	0.2849
+ (FEM × FRL)	13	-3937	0.1508	0.0001	0.2528	-2285	0.0637	0.0625	0.2849
+ (FEM × LEP)	14	-3931	0.1509	0.0460	0.2528	-2265	0.0636	0.7332	0.2849
+ (BL × FRL)	15	-3908	0.1509	0.8282	0.2528	-2248	0.0633	0.3845	0.2849
+ (BL × LEP)	16	-3884	0.1510	0.8357	0.2528	-2218	0.0634	0.9986	0.2849
+ (FRL × LEP)	17	-3858	0.1511	0.8987	0.2528	-2192	0.0634	0.8848	0.2849
Panel B: Model $K = 13$ Above With Teacher Year Effects Under Different Mixture Distributions									
MVN	19	-17298	0.1918		0.2464	-4421	0.0798		0.2836
2 Comp. GM	19	-17502			0.2462	-4593			0.2833
3 Comp. GM	19	-17467			0.2460	-4452			0.2832

Note: Results from Model Selection Tests. Top panel examines fit as model specification starts with only a single intercept (level-only) and sequentially adds variables. K represents the number of dimensions. AIC is Akaike information criterion, adj-R^2 is adjusted R-squared, LR is likelihood ratio, MSE is mean squared error. Bottom panel examines fit as we move from a multivariate normal (single component) to multi-component Gaussian mixture distributions. Model specification is fixed at $K = 19$, with the $K = 13$ specification from the top panel plus 6 year dummy variables. $A_{i(t-1)}$ is lagged test score, FEM is an indicator variable if the student is female, BL is an indicator if the student is Black, FRL is an indicator for the student's free or reduced-price lunch status, and LEP indicates if the student is designated as limited English proficiency .

specific teacher-level effect that captures the teacher's absolute advantage. Furthermore, this coefficient serves as the teacher's value-added for the baseline non-female, non-black, non-FRL, non-LEP student who has an average (zero) prior test score. The remaining coefficients represent the match effects and characterize the teacher's comparative advantage.

Panel B in Table 2 show the model selection criteria for our value-added specification under different distributional assumptions. For both subjects, AIC is maximized using a two component mixture model, providing strong evidence that teacher coefficients are not normally distributed. In addition, for both subjects, MSE continues to improve with more flexible distributions. Since the focus of our analysis is on the population distribution of the teacher coefficients we use a two component mixture in our main specification.[18]

4.3 Identification

The identification of teacher quality through value-added models has been a central focus in the literature. The primary assumption for identifying the teacher coefficients in our framework is that the components in the unobserved student-level shocks, represented by ν_{it} in Eq. (2), must be orthogonal to teacher assignment. While this assumption is similar to those discussed in traditional value-added models, the practical interpretation and implications have distinct meanings for the level effects and match effects in our model.

For the level effects, the implication for the assumption on teacher assignment is very similar to those previously addressed in the literature. Rothstein (2010) outlines that the identification of traditional level-only value-added models requires that, conditional on prior academic achievement, teacher assignments be orthogonal to underlying student ability as well as other unaccounted for educational inputs. For example, if a classroom is assigned

[18]We use MATLAB on a 2.3 GHz quad-core notebook computer. Our dataset comprises approximately 17,000 teachers, each with an average of 47 students. We defined convergence conservatively as occurring when the relative change in the log-likelihood was less than 1e-8. The level-only model with year effects included 7 teacher coefficients. Assuming a multivariate normal distribution, this model converged in 64 iterations (30 secs.). For the matching model, which included 19 teacher coefficients and assumed a multivariate normal distribution, convergence required about 700 iterations (15 mins.). Assuming a two-component mixture for the matching model, the model converged in around 1,000 iterations (60 mins.).

a teacher's aide (or any unobserved teaching tool), this can potentially bias the estimated level-effect for the teacher's value-added by adding any non-zero effect of the teacher's aide.

By contrast, the identification of match effects relies on within-classroom variation in observed student characteristics. Under the key assumption that unobserved classroom inputs (e.g., targeted supports, additional resources) either affect all students similarly or do not systematically interact with the observed characteristics z_{it} , the match effects can be identified even in the presence of such inputs. A violation arises if certain unobserved resources are directed only to a specific subgroup of students, and exposure to these resources for the target students is different based on their classroom assignment. For instance, if an aide who provides assistance exclusively to LEP students is assigned to only one classroom in a school (and aide assignment is unobserved to the econometrician), yet more than one teacher has LEP students in their classes, match-effects for all teachers would be biased.

A related concern is student sorting based on unobservable traits. If teachers known for particular comparative advantages (e.g., success with low prior-achievement students) systematically attract students who also have unobservable attributes (e.g., higher motivation or stronger parental support), then our match-effect estimates may be biased–even after controlling for students observable characteristics. Thus, while we leverage within-classroom variation to bolster identification of match effects, we rely crucially on the assumption that any sorting on unobservables is either absent or uncorrelated with the match-specific dimensions of teaching effectiveness.

To conclude, we implement in Appendix D the forecast-unbiasedness test of Chetty et al. (2014) to further validate our identification strategy and maintain comparability with the literature. A component-wise application of the test (separately to the level and match components) is inappropriate in our setting because the match component is modeled as time-invariant. As shown in Appendix D, leave-year-out predictors of time-invariant terms can be mechanically attenuated, violating the test's necessary conditions. We therefore follow the original approach and apply the test to *total* value-added, which better preserves

the test's interpretation and comparability to prior work. Across specifications, we do not reject forecast unbiasedness (see Appendix D).[19]

5 Results

This section describes the population estimates of our matching model. First, we characterize the overall importance of matching for the teachers' contribution to test scores and identify the dimensions in which matching is most prevalent. Then, using our estimates, we identify which students are most impacted by teacher assignment. The first stage estimates and the estimates from the distribution of the value-added components is reported in Appendix F.

5.1 The Overall Importance of Matching

The main estimates from our empirical model are the parameters of the population distribution of the teacher coefficients, δ. Since this is a 19-dimensional, 2-component Gaussian mixture distribution, there are over 400 parameters that describe this estimated distribution. The individual coefficients characterize the determinants of value-added. For our discussion it is useful to partition these coefficients into $\delta^{(1)}$, the first seven components representing teacher-year effects that are constant across students in a given year, and $\delta^{(2)}$, the remaining twelve coefficients related to teacher-student matching.

By partitioning the coefficients and likewise the corresponding elements of z, we can decompose the value-added when teacher j is assigned to student i from Eq. (6) into a teacher level effect and a teacher-student match effect:

$$VA_{ijt} = \underbrace{\left(\delta_j^{(1)}\right)' z_{it}^{(1)} + \left(\delta_j^{(2)}\right)' \overline{z}^{(2)}}_{\substack{\text{Level} \\ \text{Effect}}} + \underbrace{\left(\delta_j^{(2)}\right)' (z_{it}^{(2)} - \overline{z}^{(2)})}_{\substack{\text{Match} \\ \text{Effect}}}$$

[19] Finally, to further allay identification concerns, the value-added literature has also explored adjusting estimates to account for time-varying classroom shocks. These methods can also be adopted in our framework, which we discuss in Appendix E.

The first two components capture the level effect, which combines teacher-year effects and value-added for the average student. The second term describes the match effect, which is the "additional" value-added for teaching student i relative to the average student.

We use this decomposition to separate variability in value-added into the parts due to differences in teacher level effects and the match effects. The variability in match effects will be driven by a combination of the heterogeneity in the teacher population distribution of $\delta^{(2)}$, given by our estimate of $\text{Cov}(\delta)$, as well as the observed variability of the student characteristics $z^{(2)}$ in the population. Specifically, the population variance of value-added attributable to matching is given by: $\text{tr}(\text{Cov}(z^{(2)})\,\text{Cov}(\delta^{(2)}))$.

The top panel of Table 3 decomposes the total variance in value-added into the percent that is associated with level differences among teachers and the percent due to teacher-student matching.[20] Consistent with the literature, the overall variance of value-added is much larger in math than in reading, with a standard deviation of 0.24 in math and 0.15 in reading, indicating that there is substantially more variability in teachers' capacity to influence math scores compared to reading scores.

While the variance in value-added is much higher in math than reading, this is entirely driven by the level component, where the variance of value-added attributed to the match coefficients is very similar between the two subjects. Consequently, the overall importance of matching is significant. Approximately 12% of the variation in teacher value-added in math and over 25% in reading can be attributed to matching.

Another way to frame the importance of match effects is to quantify the within-teacher variation in match effects. If a teacher's within variance is near zero, they have approximately a constant value-added, suggesting no role for matching. However, if a teacher's value-added varies widely across students, they have a clear comparative advantage with certain students, suggesting an important role for matching. For teacher j, the variability in their value-

[20]The total variability of value-added is given as $\text{Var}(VA) = \bar{z}\,\text{Cov}(\delta)\bar{z}' + \text{tr}(\text{Cov}(z^{(1)})\,\text{Cov}(\delta^{(1)})) + \text{tr}(\text{Cov}(z^{(2)})\,\text{Cov}(\delta^{(2)})) + 2\,\text{tr}(\text{Cov}(z^{(1)}, z^{(2)})\,\text{Cov}(\delta^{(2)}, \delta^{(1)}))$. We define $\text{tr}(\text{Cov}(z^{(2)})\,\text{Cov}(\delta^{(2)}))$ as the variance due to matching and the rest as the variance due to the level.

Table 3: Decomposition of Teacher Contribution to Test Scores

	Math		Reading	
	Variance	(%)	Variance	(%)
Decomposition of Total Value-Added				
Total	0.0559	100.0	0.0226	100.0
Level	0.0492	88.0	0.0165	72.8
Matching	0.0067	12.0	0.0061	27.2
Decomposition of Match Effect (Partial Component Variance)†				
Lag Score	0.0040	74.1	0.0036	73.4
Female	0.0003	6.3	0.0002	3.5
Black	0.0003	6.0	0.0003	6.9
FRL	0.0005	8.6	0.0005	10.9
LEP	0.0003	5.0	0.0003	5.2

Note: This table shows the sources of the variation in teacher value-added in the population (see Footnote 20). The partial component variance for lag score, for example, is calculated by holding the variables fixed at their mean values and calculating the variance of the match component if students only varied in lag score.

added across students holding fixed their level effect is given by $(\delta_j^{(2)})' \operatorname{Cov}(z^{(2)})\delta_j^{(2)}$. Taking expectations across all teachers, the average within-teacher variability of value-added is $\mathrm{E}_\delta[(\delta^{(2)})' \operatorname{Cov}(z^{(2)})\delta^{(2)}] = \operatorname{tr}(\operatorname{Cov}(z^{(2)}) \operatorname{Cov}(\delta^{(2)}))$. Thus the variance in the match component in value-added reported in Table 3 is mathematically equivalent to the average within-teacher variation in match effects. The standard deviations of the within-teacher match effects for the average teacher in the population, $\sqrt{0.0067} = 0.082$ for math and $\sqrt{0.0061} = 0.078$ for reading, are economically large – the difference for the average teacher assigned a well- versus poorly-matched student (e.g., 1 sd above and below their mean value-added), is greater than 0.1σ test score units for reading and math.

Finally, the bottom panel of Table 3 describes which dimensions of student characteristics are most relevant for matching. For each student characteristic, we consider how much variability there would be in the match component if students only varied in that one di-

mension and all other characteristics were held fixed, at the mean value in the population.[21] Results indicate that for both math and reading, the overwhelming majority of the variability in match effects relates to prior academic achievement, accounting for around 75% of the match effects in both subjects. In contrast, matching on gender, race, FRL, or LEP plays a smaller role, with most accounting for less than 10% of the total variability of matching.

5.2 For Which Students Do Teacher Assignments Matter Most?

We can identify the students most impacted by teacher assignments by assessing the within-student variation in value-added for students with similar characteristics. Using the variation in teacher value-added to establish the importance of teacher assignment is a common practice in the literature. For example, the significantly larger variance in value-added for math compared to reading has been interpreted to mean that students are more impacted by their assigned teacher in math than in reading.

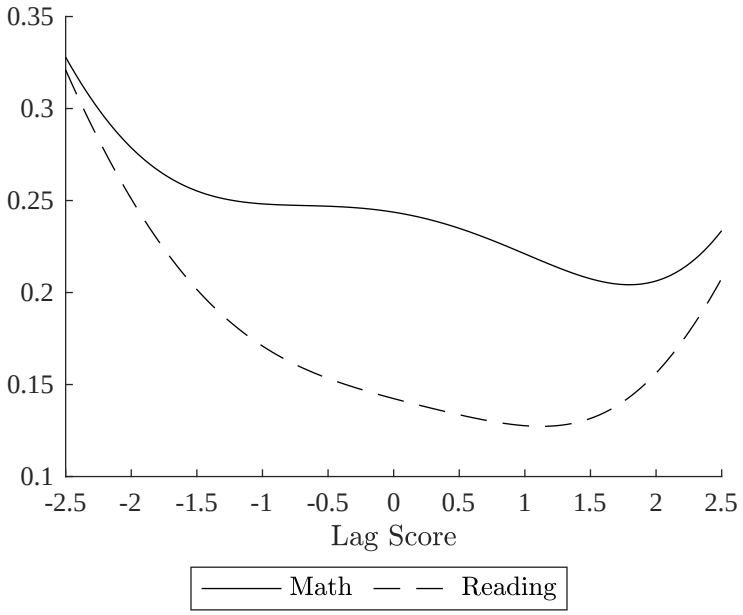
Our framework not only allows the variability in value-added to differ by subject, but it also varies at the level of the individual student. For students whose characteristics result in high variability in the teacher population for the match component of value-added, teacher assignment will significantly impact their scores. Conversely, for students with lower variability in value-added, their scores will be less affected by their assigned teacher.

We begin by discussing the variability in value-added for each of the match characteristics, holding the other characteristics fixed. Figure 1 shows the standard deviation in value-added for math and reading across different levels of lag score, holding the other match characteristics fixed.[22] This figure demonstrates that the variability in teacher value-added is significantly higher for low-ability students in both math and reading. For example, in reading, the standard deviation of value-added is nearly 33% higher for students with a lag

[21] This procedure treats each variable one at a time and does not fully take into account the interaction of all of the variables. Thus the partial component variances may sum to the total variance.

[22] Since the variability in value-added depends on the characteristics of the student, Figure 1 is constructed by sequentially assigning each value of lag score to all students in the data and reporting the average standard deviation in value-added across all students at each value.

Figure 1: Standard Deviaiton of Teacher Value-Added By Student Lagged Test Score Holding Other Variables Fixed

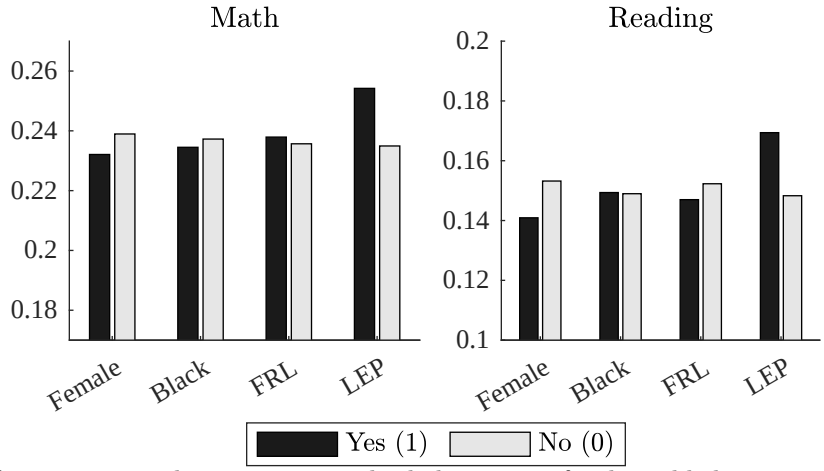


Note: This figure reports the average standard deviation of value-added across all students when we replace their lag score with the values along the x-axis. See Footnote 22 for more details.

score of -1 sd compared to those with $+1$ sd (0.171 vs. 0.128). Additionally, the figure shows that for both low-ability and high-ability students, the variability in value-added is similar between math and reading. For example, the difference in standard deviation between math and reading is 0.1 for the average student, but decreases to less than 0.028 for students who are 2 standard deviations below the mean or lower. This suggests that the large differences in value-added between math and reading previously reported in the literature mainly reflect heterogeneity in teacher value-added for students with average prior achievement, which is not necessarily true for students on the tails of the distribution.

Next, Figure 2 quantifies the within-student variation in teacher value-added across the dichotomous match variables. As shown in Table 3, these variables account for a much smaller portion of match effects, so we should expect less pronounced differences in the variation of value-added compared to those observed with lag score. First, the standard deviation in value-added is larger for males than for females in both math and reading. For reading, the standard deviation in value-added for males is 0.153, which is about 8.7% larger

Figure 2: Standard Deviation of Teacher Value-Added By Demographic Characteristics Holding Other Variables Fixed



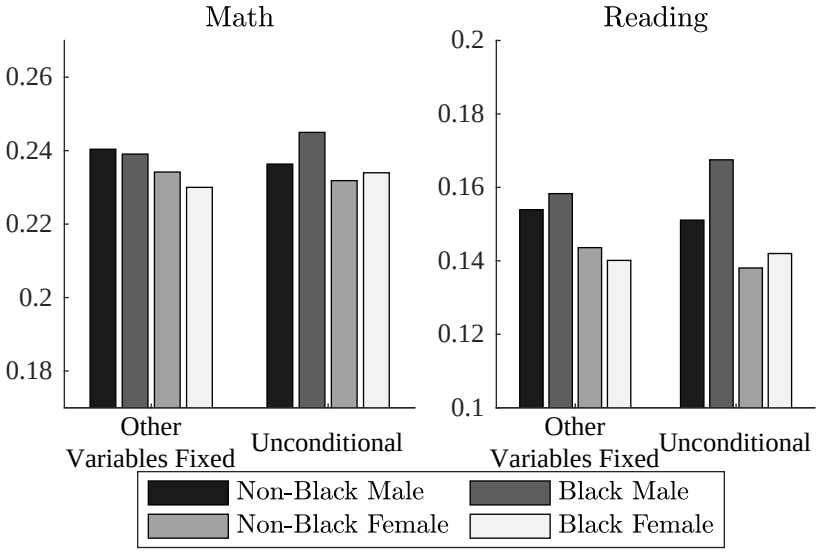
Note: This figure reports the average standard deviation of value-added across all students when assigned each value of the dichotomous variable.

than that for females. Second, although LEP students represent only a small fraction of the student population, they exhibit significant variation in teacher value-added. For example in reading the average standard deviation is about 0.17, approximately 15% larger than the standard deviation for non-LEP students. Finally, for both subjects, the variation in value-added between Black and non-Black students, as well as between FRL and non-FRL students, do not suggest significant differences in the importance of matching.

Figures 1 and 2 focus on the variability in value-added for each student characteristic while holding other characteristics fixed. However, many policy analyses may be framed at a more aggregate level. For example, rather than comparing the variability in value-added between male and female students while holding other variables fixed, a policymaker may be interested in making comparisons incorporating important differences in other dimensions of matching between the two groups of students.

To illustrate this point, Figure 3 further breaks down the variability in value-added by gender and race. It first holds other variables constant, as in the previous analysis, and then calculates the unconditional average standard deviation within group. The unconditional standard deviation accounts for group differences in lag test scores and the other matching

Figure 3: Standard Deviaiton of Teacher Value-Added By Race and Gender



variables.

Initially, holding other variables constant, non-Black males have the greatest variability in value-added for math, while Black males have the greatest variability in value-added for reading. This is consistent with the results in Figure 2, which show that males face higher variability in value-added compared to females in both subjects.

However, when calculating the unconditional standard deviation in value-added, important changes in these gaps emerge. Non-Black male students have slightly higher lag scores in math than non-Black female students, so the difference in variability in value-added shrinks. Conversely, since Black male students have significantly lower lag math scores, their variation in teacher quality increases substantially. These differences become particularly stark in reading, where the standard deviation of value-added for Black male students increases to 0.168, which is 20% larger than that for non-Black females, who have the lowest variability in teacher quality in reading.

Overall, these results demonstrate two important findings. First, when assigning students to teachers, the impact of being assigned to the 'right' versus 'wrong' teacher can be much greater for some students than for others. As we have shown that teacher assignment is especially crucial for students with lower prior academic achievement – identifying

high-quality teachers for these students can significantly improve their outcomes and help bring them up to standard. Second, because student characteristics are highly correlated, studying matching in a single dimension may not adequately capture the nature of teachers' contributions to test scores. For example, we find that matching is particularly important for Black male students, not necessarily because matching is most important by race or gender, but because it is primarily driven by lag scores, and these students tend to have lower prior achievement on average.

6 The Precision of Value-added Estimates

While the previous section focused on characterizing the population distribution of teacher value-added, we now turn to analyzing the properties of the empirical Bayes estimators for individual teachers' value-added in our data. Empirical Bayes estimators, or shrinkage estimators, are regularly used in policy analysis where predictions of individual teachers' value-added are needed.

Equation (5) shows the predictive distribution of a teacher's coefficients of value-added given the observed data, where $\mathrm{E}(\delta_j|Y_j, Z_j)$ denotes the predictive mean and $\mathrm{Var}(\delta_j|Y_j, Z_j)$ the covariance. In policy settings that center around personnel decisions, there is likely little use for predictions of the individual components of δ. Rather, what matters is how these elements can be used to form the predictive distribution of value-added for a teacher when they are matched to a particular student. For a student with characteristics z_i, assigned teacher j, the value-added is $VA_{ij} = z_i'\delta_j$ and the empirical Bayes estimator and precision of the value-added are:

$$\begin{aligned} \text{Empirical Bayes Estimator} &: \quad \mathrm{E}(VA_{ij}|z_i) = z_i'\,\mathrm{E}(\delta_j|Y_j, Z_j) \\ \text{Precision of Estimator} &: \quad \mathrm{Var}(VA_{ij}|z_i) = z_i'\,\mathrm{Var}(\delta_j|Y_j, Z_j)z_i \end{aligned}$$

Since the precision of a teacher's value-added estimate depends on the characteristics of a

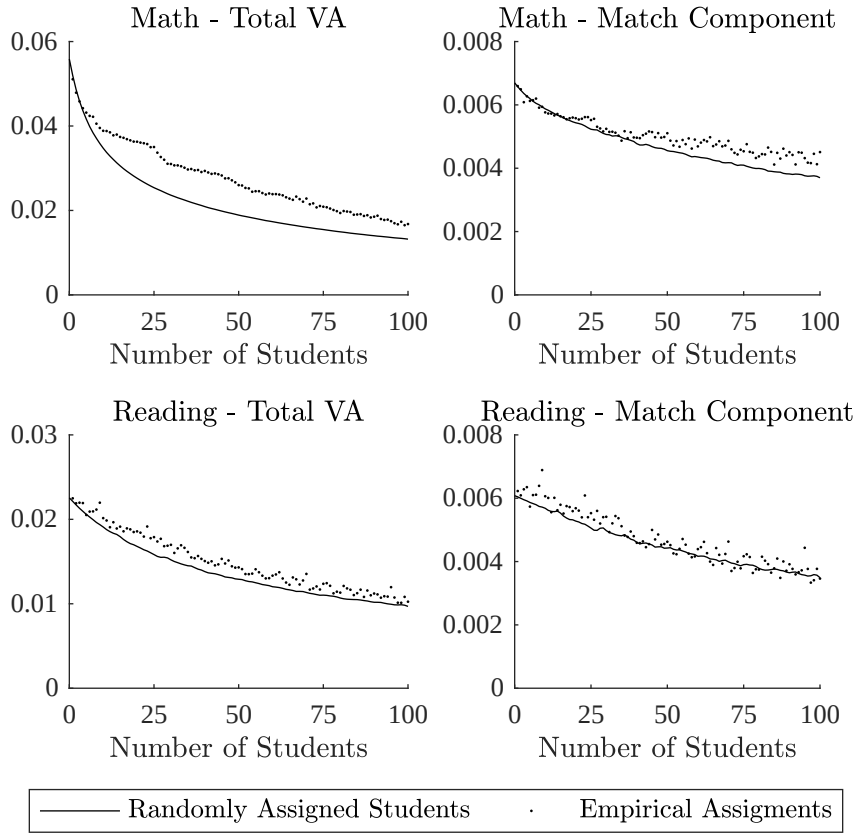
given student, a natural way to summarize the overall precision for a given teacher is by using their average precision. The average precision of a given teacher with observed test score residuals, Y_j, and student characteristics, Z_j, is $\text{Var}(VA|Y_j, Z_j) = \int (z' \text{Var}(\delta_j|Y_j, Z_j)z)h(z)dz$, where $h(z)$ representing the empirical distribution of z.

In the matching model, the precision of the value-added estimates is determined by both the number and type of students assigned to a teacher. The forecasts will be most precise for student types that the teacher frequently teaches and less precise for student types outside of this observed set. Thus, the greater the coverage of different student types assigned to a teacher, the more uniform the precision of their value-added estimates will be across students. Conversely, non-random assignment of students to teachers will result in more asymmetric precision. Focusing on the average precision across the empirical distribution of student characteristics provides a broad measure of overall precision.

Figure 4 shows how the average precision of value-added in math and reading improves as the number of observations for a given teacher increases from 0 to 100. Similar to Table 3, we can decompose the precision into the overall precision (left side of Figure 4) and the precision of the match component (right side). To understand the implications of non-random teacher-student assignments, the solid line represents the average precision with random assignment, while the scatter plot shows the average precision from the empirical assignments.

The value of the plots at zero students reflects the prior, corresponding to the population variance in value-added reported in Table 3. As additional data becomes available for a teacher, the precision improves. Starting with the precision in total value-added in math in the upper left plot of Figure 4, we observe a significant difference between the average precision that would result from random student-teacher assignments and the empirical assignments. This difference arises for two reasons. First, our data spans seven academic years, and an important component of value-added is the time-varying teacher year effects. Our measure of average precision integrates the precision of the teacher's value-added over all student types in the data covering these seven years. Empirically, for a teacher in the

Figure 4: Average Precision of Value-Added By Number of Observations



data with 20 students, these observations will generally be concentrated in a single academic year, meaning the precision of the value-added estimates will be quite accurate in this year and less so in the other six years. The line corresponding to random assignment studies the precision in the value-added if the teacher was randomly assigned 20 students from the data set, without constraining these students to be in the same academic year. The second source of non-random assignment is that, from year-to-year, teachers tend to be assigned similar types of students. This persistence in student types means that some teachers may have very little data pertaining to their value-added for certain student groups. To isolate this effect, the plots on the right show the precision in just the match component of value-added.

There are four main takeaways from Figure 4. First, for the overall precision in value-added, a teacher with around 50 student observations, or about two years of data – which is the average number of observations per teacher in our data – has a mean variance of the pre-

dictive distribution of approximately 0.0259 for math and 0.0143 for reading. This represents nearly a 50% reduction in the variance for math and around 40% for reading compared to their priors. Second, the gap in precision between random versus the empirical assignments shrinks as the number of students increases. For math teachers with 100 students, the empirical precision is 0.0168 compared to 0.0133 under random assignment. This indicates that the teacher year effects are so strongly correlated that observing 100 students concentrated in 3-4 years is nearly as informative as observing 100 students spread out across seven years.

The third takeaway is that the reduction in the uncertainty in the match component is much slower compared to the overall value-added. This is not surprising as the match component has a smaller variance than the overall variance. Furthermore, this component is much more sensitive to the types of students the teacher is assigned. Consequently, for a teacher with 100 students in math, for example, the overall uncertainty in value-added is reduced by 70%, but the uncertainty in the match component is only reduced by around 30%.

Finally, the generally higher variance in the precision of the matched component of value-added based on the empirical assignments compared to random assignments highlights that because teachers tend to be assigned to students with similar characteristics over time, this negatively impacts the precision of the matched effects, leading to less precise estimates across the full distribution of student characteristics.

Figure 4 summarizes one aspect of the precision of our estimates. It shows that for the average teacher in our data, with around 50 observations, we can reduce the uncertainty of their overall value-added by approximately 50% in math and 40% in reading. In Appendix G, we analyze other aspects of the precision of our estimates. First, we compare our matching model with the level-only model and show that the precisions of the two models are very similar. Second, we investigate how the precision of the value-added is affected by sparsely distributed students. Specifically, we look at LEP students, who comprise less than 5% of our data, and show that because of the significant overlap in student characteristics, even if

a teacher is rarely observed with LEP students, the precision of their value-added for LEP students is only slightly lower compared to teachers who have taught many LEP students.

7 Can Teacher Reallocation Lead to Student Gains?

Our analysis so far suggests that moving teachers into classrooms where they have higher match effects may lead to improvements in student test scores. The potential gains will be determined by three factors. The first is how significantly the composition of students differs across classrooms. For example, if students are very similar between two classrooms, a teacher changing classes will yield no gains due to match effects. The second factor is the heterogeneity in the match effects among the teachers being reallocated. If teachers have the same comparative advantage, then student test scores will not be impacted, even if classroom compositions are very different. Third, the gains to reallocation depend on the current allocation of teachers. If teachers are currently allocated optimally, then there are no potential gains. For example, in a scenario where we are considering swapping two teachers, there is a 50% chance they are already optimally allocated simply due to luck.

In this section, we consider the counterfactual gains in student test scores from reallocating teachers to maximize student outcomes. We examine two reallocation scenarios.[23] In the first scenario, we reallocate teachers among 4th and 5th grade classrooms within a school, which is effectively a zero-cost policy since the assignment of teachers to classrooms within a school is a routine process that occurs at the beginning of every academic year. In the second scenario, we consider reallocating teachers across schools within a district. This can generate larger gains but may entail additional costs, such as commuting expenses or teacher turnover, which are not captured by our model. For both of these reallocations, we treat the student composition of each classroom and the available teachers within a school and district as fixed as they occur in our data. Finally, since the number of students per class

[23]More ambitious policies may move teachers across districts, restructure classes, or replace current teachers. These actions could impose significant costs that may offset any gains from the reallocation.

typically varies, we focus only on gains related to the matching component, which precludes gains from assigning teachers with higher level effects to larger classrooms.

We introduce the reallocation problem in a general framework, where reallocation could occur at any number of levels. The allocation problem requires assigning S teachers, each to one of S classrooms. Let $m_{js} = 1$ indicate that teacher j is assigned to teach the students in classroom s and zero otherwise. M is a matching allocation of teachers to classrooms, where $M = \{\{m_{js}\}_{s=1}^{S}\}_{j=1}^{S}$. The allocation constraints on M are that each teacher must be assigned to one class, $\sum_{s=1}^{S} m_{js} = 1 \; \forall j$, and all classrooms must be assigned a teacher such that $\sum_{j=1}^{S} m_{js} = 1 \; \forall s$. Classroom s has n_s students, with average characteristics related to matching denoted as $\overline{z}_s^{(2)}$. For a given allocation M, the average of the component of test scores impacted by the allocation is:

$$Y(M|\delta_1, \delta_2, \ldots, \delta_S) = \frac{1}{\sum_{s=1}^{S} n_s} \sum_{s=1}^{S} \sum_{j=1}^{S} m_{js} \times n_s \times \left(\overline{z}_s^{(2)\prime} \delta_j^{(2)} \right)$$

Denote $M^* = \text{argmax}_M \, Y(M|\delta_1, \delta_2, \ldots, \delta_S)$ as the optimal allocation of teachers that maximize average test scores. The problem is not as simple as assigning teachers to their highest match classroom because the opportunity cost of assigning a teacher to a classroom is that another teacher cannot be assigned to that class. Thus the optimal assignment takes into account a teacher's comparative advantage relative to the other teachers in the school.

In our counterfactual analysis we would like to calculate the gains in student test scores that can be produced by optimally allocating teachers compared to the status quo. Given the observed classroom assignment, M^o, the gains in test scores from the optimal assignment of teachers to classrooms is given by $Y(M^*) - Y(M^o)$.[24] Because each teacher's match coefficients are not precisely known, the gains need to estimated using the posterior densities

[24]We asume that match coefficients remain fixed as assignments change. That is, all changes in a teacher's value-added is through differences in the composition of z, while δ_j stays fixed. Aucejo et al. (2019) finds little evidence that teachers adapt their teaching practices when changing classrooms. Thus the likelihood that teachers change their behavior when assigned to a new class would not appear to be a large concern.

in Eq. (5) for each teacher. Specially, we seek an estimate of the expectation of the gains:

$$\mathrm{E}\left[Y(M^*|\delta_1, \delta_2, \ldots, \delta_S) - Y(M^o|\delta_1, \delta_2, \ldots, \delta_S)\right] \tag{7}$$

To produce an unbiased estimate of Eq. (7), we integrate over the full posterior distributions in Eq. (5). Since $Y(\cdot)$ is a non-linear function, we use numerical integration.[25]

For each level of reallocation, we draw from each teacher's posterior distribution of their match coefficients and calculate the average gain in test scores given an optimal reallocation. We repeat this process 1,000 times for each level of reallocation, and the expected gains are the average from these 1,000 simulations.

Columns (1) and (2) in Table 4 show the average gains for math and reading when teachers are optimally reallocated within schools.[26] The overall gain in average test scores is quite large, increasing by 0.0227 standard deviations in math and 0.0192 in reading. To put these numbers in context, Aucejo and Romano (2016) found that adding 10 days of instruction to the school year would improve math scores by 0.017 standard deviations and reading scores by 0.008 standard deviations. Our estimated gains from reallocating teachers within schools are larger than these estimates and can be implemented with nearly zero costs, whereas increasing the number of instructional days could be very costly.

As Section 5 demonstrates that certain students are more impacted by teacher assignments than others, Table 4 shows how the overall gains are dispersed by prior academic achievement and by race and gender. Consistent with Figure 1, students in the lower and upper tails of the prior achievement distribution experience the largest increases in test scores from reallocating teachers. The gains are especially pronounced in reading for students in the bottom-third of the lag score distribution, who experience an increase of 0.0358 standard deviations. Following Figure 3 that showed Black male students were the most impacted by

[25]Swapping the maximization with the expectation in Eq.(7), such that the equation is $Y(M^*|\,\mathrm{E}(\delta)) - Y(M^o|\,\mathrm{E}(\delta))$, where $\mathrm{E}(\delta)$ are the empirical Bayes estimators, downward biases estimates of the gains from reallocation as the maximum of the expected value is always less than the expected value of the maximum.

[26]The simulations calculate gains in reading and math separately, meaning two teachers, one for reading and one for math, can be assigned to the same classroom.

Table 4: Counterfactual Average Test Score Gains From Reallocation

	Optimally Allocated Within School		Optimally Allocated Within District	
	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
Overall	0.0227	0.0192	0.0535	0.0466
Gains By Lag Score Percentile				
Bottom One-Third	0.0303	0.0358	0.0696	0.0840
Middle One-Third	0.0083	0.0067	0.0224	0.0186
Top One-Third	0.0297	0.0187	0.0685	0.0453
Gains By Race/Gender				
Non-Black Males	0.0219	0.0190	0.0503	0.0435
Black Males	0.0288	0.0250	0.0715	0.0653
Non-Black Females	0.0215	0.0168	0.0500	0.0404
Black Females	0.0228	0.0208	0.0547	0.0548

Note: Counterfactual average test score gains (in standard deviation units) from reallocating teachers within schools and across schools (within the same district). Teachers are reassigned to classrooms that correspond to their original grade level, and original classroom composition is maintained.

teacher assignment, these students benefit the most from reallocating teachers, with average increases in test scores of 0.0288 standard deviations in math and 0.0250 in reading.

Finally, Columns (3) and (4) show the gains from reallocating teachers across schools within a district. The gains are about twice as large compared to only reallocating teachers within schools. This is not surprising, as there is a greater role for comparative advantage when there are more diverse classrooms and heterogeneity among teachers.[27]

Overall, these counterfactual simulations highlight the importance of matching and indicate that optimal reallocations of teachers could lead to meaningful improvements in students' performance, even in scenarios where teacher reassignments are highly restricted.

[27]Laverde et al. (2023) also shows that potential gains from within-district teacher reallocations are larger than those from within-school reallocations.

8 Teacher Assignment and the Ranking of Teachers

A current debate among policymakers and scholars is how to better use value-added estimates to make personnel decisions. For example, Chetty et al. (2014) and Hanushek (2009) study the potential benefits of replacing teachers at the bottom 5% of the value-added distribution. Springer et al. (2010) examines a pilot performance pay program in Tennessee that offered bonuses for generating value-added beyond ambitious historic thresholds. We extend this literature by accounting for how match effects may impact teacher performance and thus, their ranking relative to other teachers (or distance to a defined threshold).

8.1 Sensitivity of Teacher Ranking to Classroom Assignment

To study how sensitive teacher rankings are to classroom assignment, we calculated each teacher's best linear unbiased predictors of their match effects and ranked teachers based on their average value-added for their assigned classroom $\overline{VA}_{js^o} = \overline{z}'_{s^o}\hat{\delta}_j^{BLUP}$, where s^o corresponds to teacher j's original classroom assignment. Holding this ranking of value-added as fixed, for each teacher we calculated the change in their percentile ranking if they were assigned to the classroom in their school or district in which they have the highest average match. The results from this analysis are summarized in Table 5.

Table 5 shows that around 17-18% of teachers are already assigned to their best-matched classroom in their school. Approximately 17.2% of teachers would experience at least a 5 percentage point increase in their reading percentile rank if they were assigned to their highest matched classroom at the school. For math, the potential gains are less pronounced, with around 8% of teachers experiencing more than a 5 percentage point increase in their percentile rank. These gains are amplified when we look at the tails of the value-added distribution, where most of the policy discussions have focused. The bottom 5th percentile teacher in the math ranking has a value-added of -0.3014, while for reading, it is -0.1565. Among reading teachers who originally ranked in the bottom 5% based on their value-added

Table 5: Impact of Teacher Reallocation on Rankings When Teachers Are Assigned Best Match in School Compared to Original Assignment

	Comparison to Best Match in School		Comparison to Best Match in District	
	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
Share with No Change	0.177	0.178	0.018	0.017
Share with Change in Rank >5%	0.080	0.172	0.361	0.541
Share with Change in Rank >10%	0.016	0.043	0.115	0.249
Share No Longer in Bottom 5% Rank	0.210	0.292	0.502	0.654

Note: This table characterizes changes in teachers' ranking when teachers are assigned to their best possible classroom in the school or district. School denotes only grades 4 and 5. The baseline ranking is estimated based on the average value-added of the teacher when considering only her average originally-assigned students. Only classes with between 5 and 40 students were used.

scores, 29.2% would escape the bottom 5% if they were reassigned to their highest-matched classroom in their school. Furthermore, if these teachers were reassigned to a different classroom anywhere within the district, 65.4% would no longer be in the bottom 5%.[28]

As demonstrated in Table 5, teacher rankings can be highly sensitive to classroom assignments. Understanding and accounting for these matching effects can not only improve education production but also help administrators to make good personnel decisions, by providing an estimate of the value-added of teachers in other classrooms where they may be assigned in the future. The purpose of this exercise is not to suggest a new ranking system, but to demonstrate the degree to which a lucky (or unlucky) class assignment can impact current teacher rankings. Appendix H proposes a transparent approach to rank teachers that avoids the idiosyncrasies that may arise from realized classroom assignments.

9 Conclusion

This study explores the importance of student-teacher matching using a novel, highly flexible framework for estimating multivariate value-added models. The estimation framework,

[28]For math, the impacts are slightly attenuated, 21% and 50% at the school and district level, respectively

based on maximum likelihood, relies on weaker distributional assumptions for value-added. Additionally, we discuss model selection for both population fit and predictive forecasting of value-added for conditional student-teacher assignments.

We estimate our matching model using administrative data from public schools in North Carolina. Student-teacher match effects are significant and comparable between math and reading, with the average teacher's effectiveness differing by 0.1σ test score units between poorly-matched and well-matched students. Match effects are especially important for low-achieving students in reading, where the standard deviation of teacher value-added is 33% higher than for high-achieving students. This suggests that identifying high-quality teachers can be particularly beneficial for underperforming students.

Using our results, we conduct a counterfactual analysis to estimate the test score gains from optimally assigning teachers to classrooms. We find significant gains from reallocating teachers within schools, with overall test score improvements of about 0.02 sd. Larger gains of approximately 0.05 sd could be achieved by reallocating teachers across schools within a district. The test score gains are particularly substantial for traditionally disadvantaged students. For example, reallocating teachers within districts could increase average math scores for Black male students by 0.07 sd.

References

A. Abdulkadiroğlu, P. A. Pathak, J. Schellenberg, and C. R. Walters. Do parents value school effectiveness? *American Economic Review*, 110(5):1502–39, 2020.

T. Ahn, E. Aucejo, and J. James. Revisiting tests of forecast unbiasedness in teacher value-added models. Working paper, California Polytechnic University, 2025.

E. M. Aucejo and T. F. Romano. Assessing the effect of school days and absences on test score performance. *Economics of Education Review*, 55:70–87, 2016.

E. M. Aucejo, P. Coate, J. C. Fruehwirth, S. Kelly, and Z. Mozenter. Match effects in the teacher labor market: Teacher effectiveness and classroom composition. 2019.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

M. D. Bates, M. Dinerstein, A. C. Johnston, and I. Sorkin. Teacher labor market equilibrium and student achievement. Technical report, National Bureau of Economic Research, 2022.

B. Biasi, C. Fu, and J. Stromme. Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage. Technical report, National Bureau of Economic Research, 2021.

J. Braun, L. Held, and B. Ledergerber. Predictive cross-validation for the choice of linear mixed-effects models with application to data from the swiss hiv cohort study. *Biometrics*, 68(1):53–61, 2012.

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

S. Condie, L. Lefgren, and D. Sims. Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, 40:76–92, 2014.

W. Delgado. Heterogeneous teacher effects, comparative advantage, and match quality. Technical report, Boston University, 2021.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

S. Gershenson, C. Hart, J. Hyman, C. Lindsay, and N. W. Papageorge. The long-run impacts of same-race teachers. Technical report, National Bureau of Economic Research, 2018.

M. Gilraine and N. Pope. Making teaching last: Long- and short-run value-added. Technical report, Working Paper, 2020.

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

J. Gong, Y. Lu, and H. Song. The effect of teacher gender on students' academic and noncognitive outcomes. *Journal of Labor Economics*, 36(3):743–778, 2018.

B. S. Graham, G. Ridder, P. Thiemann, and G. Zamarro. Teacher-to-classroom assignment and student achievement. *arXiv preprint arXiv:2007.02653*, 2020.

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

C. K. Jackson. Match quality, worker productivity, and worker mobility: Direct evidence

from teachers. *Review of Economics and Statistics*, 95(4):1096–1116, 2013.

M. Jamshidian and R. I. Jennrich. Acceleration of the em algorithm by using quasi-newton methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59 (3):569–587, 1997.

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

M. Laverde, E. Mykerezi, A. Sojourner, and S. Aradhya. Gains from reassignment: Evidence from a two-sided teacher market'. Technical report, 2023.

V. Lavy. What makes an effective teacher? quasi-experimental evidence. *CESifo Economic Studies*, 2015.

L. Lusher, D. Campbell, and S. Carrell. Tas like me: Racial interactions between graduate teaching assistants and undergraduates. *Journal of Public Economics*, 159:203–224, 2018.

E. Marshall and D. Spiegelhalter. Approximate cross-validatory predictive checks in disease mapping models. *Statistics in medicine*, 22(10):1649–1660, 2003.

G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

N. Petek and N. G. Pope. Learning by doing vs. learning about match quality: Can we tell them apart? *Working Paper, University of Maryland*, 2021.

J. Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214, 2010.

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

M. Springer, D. Ballou, L. Hamilton, V. Le, J. Lockwood, Mc-Cafrey, M. D., Pepper, and B. Stecher. Teacher pay for performance: Experimental evidence from the project on incentives in teaching. Technical report, National Center on Performance Incentives at Vanderbilt University, 2010.

M. P. Steinberg and R. Garrett. Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, 38(2):293–317, 2016.

D. O. Stram and J. W. Lee. Variance components testing in the longitudinal mixed effects model. *Biometrics*, pages 1171–1177, 1994.

F. Vaida and S. Blanchard. Conditional akaike information for mixed-effects models. *Biometrika*, pages 351–370, 2005.

Appendix For Online Publication

A Analytic Sample Selection

Our sample is constructed in the following manner:

Table A1: Construction of Analytic Sample

#	**Description**	**Observations**	**% of Sample**
1	Start with a raw data of 4th and 5th grade student/year observations.	1,593,605	100.0%
2	Drop students who cannot be linked to both math and reading teachers.	-235,606	-14.8%
3	Drop students who have different math and reading teachers (non-self-contained classes).	-378,425	-23.7%
4	Drop students who are missing both reading and math scores.	-86,685	-5.4%
5	Drop students with missing RHS variables	-77,826	-4.9%
6	Drop teachers (and their students) who switch schools during the academic year, have more than 50 students or less than 5 students in their class	-10,184	-0.6%
7	Obtain our sample	804,879	50.5%

B Estimation Details

Given an unbiased estimate $\hat{\beta}$ for the non-valued-added component of test scores, for each teacher j we construct their $n_j \times 1$ vector of residuals $Y_j = A_j - X_j\hat{\beta} = Z_j\delta_j + \nu_j$, which from Eq. (3) follows $Y_j|\delta_j \sim N(Z_j\delta_j, \sigma^2 I_{n_j})$, where I_{n_j} is the $n_j \times n_j$ identity matrix. Given

the normality of Y_j we could estimate the distribution of the match coefficients directly using a random coefficients framework. However, the dimension of Y_j can be quite large in administrative data settings, leading to a computationally intensive optimization problem. Instead, we project the residuals onto the student match characteristics, which reduces the dimensionality of the data from n_j to K.

Projecting the residuals onto the match characteristics with least squares requires Z_j to be full rank. However, for some teachers, Z_j might not be full rank so, projection can only be done using a subset of the columns of Z_j. A lack of full rank will arise for a number of reasons. First, it may arise mechanically, for example in a model where the teacher's value-added varies over time, but not all teachers are observed in all years. Second, if some of the attributes in z are disproportionately allocated among teachers, for example, if matching depends on student LEP status but some teachers either never or always teach LEP students, then this would cause Z_j to not be full rank. Finally, Z_j would not be full rank if $n_j < K$, that is the number of observations for teacher j is less than the dimension of δ. A lack of full rank is not limiting. For each teacher, we define Z_j^* as any set of linearly independent columns of Z_j, such that $\text{rank}(Z_j^*) = \text{rank}(Z_j)$. If $\text{rank}(Z_j) = K$ then by definition $Z_j^* = Z_j$. Multiplying Y_j by $(Z_j^{*'} Z_j^*)^{-1} Z_j^{*'}$ creates for each teacher the least squares estimates that are distributed:

$$\hat{\delta}_j^{LS} \sim N(W_j \delta_j, \sigma^2 (Z_j^{*'} Z_j^*)^{-1})$$

The matrix W_j is a $\text{rank}(Z_j) \times K$ matrix that maps the least squares coefficients to the unobserved teacher coefficients δ_j.[29] If Z_j is full rank, then W_j is the identity matrix, providing a one-to-one mapping of the least squares coefficients to δ_j.[30] The least squares

[29] The matrix $W_j = (Z_j^{*'} Z_j^*)^{-1} Z_j^{*'} Z_j$.

[30] If Z_j is not full rank then either W_j provides a one-to-one mapping from the least squares estimates to a subset of δ_j or the least squares estimates will map to some linear combination of δ_j. Because of this second point, caution must be exercised in comparing the least squares estimates across teachers, where differences in W_j will potentially imply different interpretations of the estimates.

estimates act as sufficient statistics for the teacher residuals but reduce the dimension of the teacher data from n_j to $\text{rank}(Z_j)$, making the estimation of the model more tractable.

The log-likelihood function is constructed based on the sampling distribution of the least squares estimates. We estimate σ^2 separately from Ψ and define $\hat{\Sigma}_j = \hat{\sigma}^2(Z_j^{*'}Z_j^*)^{-1}$ as the variance-covariance matrix of the least squares estimates. Given $\hat{\nu}_j = Y_j - Z_j^{*'}\hat{\delta}_j^{LS}$, an estimate of the component of the teacher's residuals that cannot be attributed to matching, an unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{J} \hat{\nu}_j'\hat{\nu}_j}{\sum_{j=1}^{J}(n_j - \text{rank}(Z_j^*))} \tag{8}$$

Let $h(\hat{\delta}_j^{LS}|W_j\delta_j, \hat{\Sigma}_j)$ denote the density function of the least squares estimate and its relationship to the true underlying teacher coefficients δ_j. We estimate the population parameters of the distribution of the teacher coefficients by maximizing an integrated likelihood function, which accounts for the sampling error in the least squares estimates:

$$\widehat{\Psi} = \underset{\Psi}{\text{argmax}} \sum_{j=1}^{J} \ln\left(\int h(\hat{\delta}_j^{LS}|W_j\delta, \hat{\Sigma}_j) f(\delta|\Psi) d\delta\right) \tag{9}$$

Where $f(\delta|\Psi) = \sum_{c=1}^{C} \pi_c \phi(\delta|\gamma_c, \Delta_c)$ and $\Psi = \{\pi_c, \gamma_c, \Delta_c\}_{c=1}^{C}$.

B.1 Optimization of the Log-Likelihood with the EM Algorithm

Our estimator of the population distribution of the value-added coefficients assumes that this distribution can be approximated by a C component Gaussian mixture, such that $f(\delta|\Psi) = \sum_{c=1}^{C} \pi_c \phi(\delta|\gamma_c, \Delta_c)$, where π is the share parameter for each component and $\phi(\delta|\gamma_c, \Delta_c)$ is the probability density function of a multivariate normal distribution with component-specific mean γ_c and covariance Δ_c. Estimation of the Gaussian mixture model will be based on the least squares estimates $\hat{\delta}_j^{LS}$, which have a sampling distribution, $h(\hat{\delta}_j^{LS}|W_j\delta_j, \hat{\Sigma}_j)$. Combining these individual distributions with the population distribution for δ, the log-

likelihood represented in Eq. (9) takes the form:

$$LL(\Psi) = \sum_{j=1}^{J} \ln \left(\sum_{c=1}^{C} \pi_c \phi \left(\hat{\delta}_j^{LS} | W_j \gamma_c, \hat{\Sigma}_j + W_j \Delta_c W_j' \right) \right) \tag{10}$$

Where $\Psi = \{\pi_c, \gamma_c, \Delta_c\}_{c=1}^{C}$ is the vector of parameters to be estimated and again ϕ represents the probability density function of a multivariate normal distribution, except now the covariance matrix is $\hat{\Sigma}_j + W_j \Delta_c W_j'$ rather than simply Δ_c.

The parameters of the Gaussian mixture model are estimated by maximizing the log-likelihood in Eq. (10), which we optimize with the expectation-maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm is an established method for estimating Gaussian mixture models. However, its conventional implementation cannot be deployed in our setting because the data on which estimation is based is measured with error. Instead, we use the following modified iterative algorithm.

- Initialize with starting values $\Psi^{(0)} = \{\pi_c^{(0)}, \gamma_c^{(0)}, \Delta_c^{(0)}\}_{c=1}^{C}$
- Repeat the following iteration until converged (i.e., $||\Psi^{(m+1)} - \Psi^{(m)}||_\infty < \kappa$)

  Expectation Step: Given $\Psi^{(m)}$, for each j and component c, compute

$$q_{jc}^{(m)} = \frac{\pi_c^{(m)} \phi(\hat{\delta}_j^{LS} | W_j \gamma_c^{(m)}, W_j \Delta_c^{(m)} W_j' + \hat{\Sigma}_j)}{\sum_{c'=1}^{C} \pi_{c'}^{(m)} \phi(\hat{\delta}_j^{LS} | W_j \gamma_{c'}^{(m)}, W_j \Delta_{c'}^{(m)} W_j' + \hat{\Sigma}_j)}$$

$$E_{jc}^{(m)} = \gamma_c^{(m)} + \Delta_c^{(m)} W_j' \left(W_j \Delta_c^{(m)} W_j' + \hat{\Sigma}_j \right)^{-1} \left(\hat{\delta}_j^{LS} - W_j \gamma_c^{(m)} \right)$$

$$U_{jc}^{(m)} = \Delta_c^{(m)} - \Delta_c^{(m)} W_j' \left(W_j \Delta_c^{(m)} W_j' + \hat{\Sigma}_j \right)^{-1} W_j \Delta_c^{(m)} + \left(E_{jc}^{(m)} \right) \left(E_{jc}^{(m)} \right)'$$

  Maximization Step: Update parameters $\Psi^{(m+1)}$. For each component c com-

pute

$$\pi_c^{(m+1)} = \frac{1}{J}\sum_{j=1}^{J} q_{jc}^{(m)}$$

$$\gamma_c^{(m+1)} = \left(\sum_{j=1}^{J} q_{jc}^{(m)} \left(E_{jc}^{(m)}\right)\right) / \left(\sum_{j=1}^{J} q_{jc}^{(m)}\right)$$

$$\Delta_c^{(m+1)} = \left(\sum_{j=1}^{J} q_{jc}^{(m)} \left(U_{jc}^{(m)}\right)\right) / \left(\sum_{j=1}^{J} q_{jc}^{(m)}\right) - \left(\gamma_c^{(m+1)}\right)\left(\gamma_c^{(m+1)}\right)'$$

Although the EM algorithm outlined above is easy to implement and globally convergent, its iterations are often slow to converge. To accelerate convergence, we modify the EM step size using the quasi-Newton method outlined in Jamshidian and Jennrich (1997), which drastically increases the rate of convergence. To estimate the standard errors, we use an estimate of the observed information matrix of the log-likelihood in Eq. (10). Letting $\ln L_j(\Psi)$ denote teacher j's contribution to the log-likelihood function, where Eq. (10) can be written as $LL(\Psi) = \sum_{j=1}^{J} \ln L_j(\Psi)$. We can denote teacher j's contribution to the score vector as $s_j(\Psi) = \partial \ln L_j(\Psi)/\partial\Psi$. Then as discussed in McLachlan and Krishnan (2007) the empirical observed information matrix is given by

$$\mathcal{I}_e(\widehat{\Psi}) = \sum_{j=1}^{J} s_j(\widehat{\Psi}) s_j(\widehat{\Psi})'$$

C Joint Estimation of Multiple Outcomes

The estimator discussed in Appendix B uses the teacher residuals Y_j to estimate the joint distribution of the multivariate vector of teacher coefficients δ_j. In many studies, the teacher's contribution to student outcomes is analyzed on multiple outcomes (e.g., math and reading performance). These analyses are often conducted separately. However, a natural extension of our multivariate value-added framework is to jointly model these multiple outcomes in a combined analysis. There are two benefits to this approach. First, researchers may be interested in the correlation of the teacher's outcome-specific coefficients across the outcomes, for

example, to study whether teacher comparative advantages are transferable across subjects. The joint model will produce direct estimates of these parameters. Second, if outcome-specific teacher coefficients are correlated, then using the data from all of the outcomes to form the posterior distributions can lead to more precise predictions of the teacher coefficients compared to using the data from the outcomes separately.

This joint analysis requires very little modification to the estimator discussed in Appendix B, only a re-definition of the variables. For exposition, assume two sets of residuals $Y_j^{(1)}$ and $Y_j^{(2)}$ on outcomes (1) and (2) that are informative about two sets of teacher coefficients $\delta_j^{(1)}$ and $\delta_j^{(2)}$.[31] The joint distribution of these residuals analogous to Eq. (3) is:

$$\begin{bmatrix} Y_j^{(1)} \\ Y_j^{(2)} \end{bmatrix} \sim N\left(\begin{bmatrix} Z_j^{(1)} & 0 \\ 0 & Z_j^{(2)} \end{bmatrix} \begin{bmatrix} \delta_j^{(1)} \\ \delta_j^{(2)} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_2^2 \end{bmatrix} \otimes I_{n_j} \right) \tag{11}$$

$$Y_j \sim N\left(Z_j \delta_j, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \otimes I_{n_j} \right)$$

Thus defining Y by concatenating the residuals from all of the outcomes and defining Z as a block diagonal of the outcome-specific covariates provides the inputs needed to estimate the joint distribution of the teacher coefficients across all of the outcomes. The only new feature that arises in the multiple outcome framework is σ_{12}, which is the covariance in the idiosyncratic component of the residual across outcomes. This accounts for the fact that these outcomes often correspond to the same observational unit, for example, math and reading scores from the same student, which should not be treated as independent observations. We propose an unbiased estimator for this parameter below.

[31] This approach can be extended to any number of outcomes.

C.1 Implementation of Joint Estimation of Multiple Outcomes

The first stage for each outcome is estimated separately, but the distribution of all of the match effects is estimated jointly in the second stage. Assume two outcomes are to be estimated jointly. Let $Y_j^{(1)}$ and $Y_j^{(2)}$ on outcomes (1) and (2) reflect two sets of teacher coefficients $\delta_j^{(1)}$ and $\delta_j^{(2)}$. The joint distribution of these residuals analogous to Eq. (3) is:

$$\begin{bmatrix} Y_j^{(1)} \\ Y_j^{(2)} \end{bmatrix} \sim N\left(\begin{bmatrix} Z_j^{(1)} & 0 \\ 0 & Z_j^{(2)} \end{bmatrix} \begin{bmatrix} \delta_j^{(1)} \\ \delta_j^{(2)} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_2^2 \end{bmatrix} \otimes I_{n_j} \right) \tag{12}$$

$$Y_j \sim N\left(Z_j\delta_j, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \otimes I_{n_j} \right)$$

Unbiased estimates of σ_1^2 and σ_2^2 can be attained using the same estimator in Eq. (8). However, we need an unbiased estimate of σ_{12}. Given that $\sigma_{12} = \mathrm{E}(\nu_{it}^{(1)}\nu_{it}^{(2)})$, where $\nu_{it}^{(1)}$ and $\nu_{it}^{(2)}$ is student i's residual for outcomes 1 and 2, we will form this estimator based on the estimated residuals, $\hat{\nu}_{it}^{(1)}$ and $\hat{\nu}_{it}^{(2)}$, where for example:

$$\hat{\nu}_{it}^{(1)} = A_{it}^{(1)} - x_{it}\hat{\beta}^{(1)} - z_{it}\hat{\delta}_j^{(1)_{LS}}$$

To characterize the estimator, let $Y^{(1)}$ be the stacked vector of all of the teacher residuals formed from the first stage estimates for outcome 1, $Z^{(1)}$ be a block diagonal matrix of student attributes, where the diagonal elements are formed from $Z_j^{(1)}$, and $\delta^{(1)}$ a stacked vector of all of the teacher coefficients δ_j, then the entire vector of residuals can be characterized as:

$$Y^{(1)} = Z^{(1)}\delta^{(1)} + \nu^{(1)}$$

Multiplying both sides by the residual maker for $Z^{(1)}$, denoted $\mathrm{M}_{Z^{(1)}}$, gives:

$$\mathrm{M}_{Z^{(1)}}Y^{(1)} = \hat{\nu}^{(1)} = \mathrm{M}_{Z^{(1)}}\nu^{(1)}$$

Because not all students will have observed values for all outcomes, for example, some students might have observed math scores but no observed reading scores, the covariance can only be estimated from observations that have non-missing values for both outcomes. Let $D^{(1)}$ be a matrix that maps which outcomes have non-missing values for both outcomes, for example, if there are 900,000 students total in the sample, but only 823,000 have non-missing values for both outcomes, then $D^{(1)}$ is $823,000 \times 900,000$. Then we can estimate σ_{12} with the appropriate degrees of freedom adjustment as

$$\begin{aligned}
(D^{(1)}\hat{\nu}^{(1)})'(D^{(2)}\hat{\nu}^{(2)}) &= \mathrm{E}\,\mathrm{tr}\left((D^{(1)}\mathrm{M}_{Z^{(1)}}\nu^{(1)})'(D^{(2)}\mathrm{M}_{Z^{(2)}}\nu^{(2)})\right) \\
&= \mathrm{E}\,\mathrm{tr}\left((D^{(1)}\mathrm{M}_{Z^{(1)}})'(D^{(2)}\mathrm{M}_{Z^{(2)}})(\nu^{(2)})(\nu^{(1)})'\right) \\
&= \mathrm{tr}\left((D^{(1)}\mathrm{M}_{Z^{(1)}})'(D^{(2)}\mathrm{M}_{Z^{(2)}})\right)\sigma_{12} \\
\sigma_{12} &= \frac{(D^{(1)}\hat{\nu}^{(1)})'(D^{(2)}\hat{\nu}^{(2)})}{\mathrm{tr}\left((D^{(1)}\mathrm{M}_{Z^{(1)}})'(D^{(2)}\mathrm{M}_{Z^{(2)}})\right)} \qquad (13)
\end{aligned}$$

The modeling of multiple outcomes can be easily extended beyond two by modifying the joint distribution in Eq. (12). In this case, the only additional step is the estimate the additional pairwise covariances, for example, σ_{13} and σ_{23} in a model with three outcomes. These can be estimated in a similar manner to Eq. (13).

D Chetty et al. (2014) Forecast Unbiased Test

In this appendix, we implement the Chetty et al. (2014) quasi-experimental validation within our setting. This test relies on leave-year-out value-added predictors because it forecasts year-over-year changes in average test scores at the school-by-grade level; consequently, leave-two-years-out predictors are required to avoid mechanical correlation between predictors and residuals. Although our model shares key features with Chetty et al. (2014)–notably time-varying level effects that might suggest a straightforward adaptation–the inclusion of match effects is a substantive departure. The main complication is the behavior of leave-year-out

predictors for time-invariant components, which Ahn et al. (2025) has shown can induce mechanical attenuation bias in some settings.

Therefore, we first assess the forecast validity of the leave-year-out predictors by estimating student-level regressions, in the spirit of Table 3 in Chetty et al. (2014). This implies decomposing test scores into their main components and estimate different coefficients λ_n for each of them:

$$A_{it} = \lambda_1 \underbrace{[x'_{it}\beta]}_{\substack{\text{Deterministic} \\ \text{Component}}} + \lambda_2 \underbrace{\left[\left(\hat{\delta}_j^{(1)}\right)' z_{it}^{(1)}\right]}_{\substack{\text{Level} \\ \text{Component} \\ \text{of VA}}} + \lambda_3 \underbrace{\left[\left(\hat{\delta}_j^{(2)}\right)' z_{it}^{(2)}\right]}_{\substack{\text{Match} \\ \text{Component} \\ \text{of VA}}} + \nu_{it} \tag{14}$$

Table D1 reports student-level regressions of Eq. (14) estimated with leave-year-out predictors for both a CFR year-effects-only specification and our matching model. Column 1 presents the CFR-style specification with only year effects–directly analogous to Table 3 in Chetty et al. (2014)–and yields an estimate of $\lambda_2 = 0.9802$, very close to the benchmark value of 1. Column 2 reports results for the matching model, which decomposes value-added into level and match components. The estimated coefficient on λ_3 is 0.6737 –instead of the (a priori) expected coefficient of 1– indicating substantial attenuation toward zero. Notably, as discussed by Ahn et al. (2025), this attenuation does not reflect non-random assignment; it arises mechanically from using leave-year-out predictors for time-invariant components. Therefore, the main takeaway from Table D1 is that leave-year-out predictors are invalid for time-invariant match effects. As a result, the test's central identifying condition fails, which suggests caution in attempting to apply the test separately to the level and match components of our model.

Despite this limitation, to facilitate comparison with prior studies we implement the test on total value added (the sum of level and match effects), allowing direct comparison with the existing literature while avoiding component-specific interpretations. Because match

Table D1: Student-Level CFR Regressions

	CFR, Year-Effects Only	Matching Model w/ Year Effects
	(1)	(2)
$\hat{\lambda}_1$	1.0067	0.9998
	(0.0005)	(0.0005)
$\hat{\lambda}_2$	0.9802	0.9273
	(0.0036)	(0.0035)
$\hat{\lambda}_3$	–	0.6737
		(0.0094)
Obs.	1,606,907	1,606,907

Note: Student-level regressions of Eq. (14) estimated with leave-year-out predictors for both a CFR year-effects-only specification (column 1) and our matching model (column 2). Regressions combine math and reading scores.

effects constitute about 12% of value added in math and 27% in reading, the influence of the problematic time-invariant component on the total predictor should be limited. To test forecast unbiasedness of total value added, we define the average test score and average leave-two-year-out value added estimate in school s, grade g, and year t as $\overline{A}_{sgt} = (1/n_{sgt}) \sum A_{it}$ and $\overline{VA}_{sgt} = (1/n_{sgt}) \sum VA_{ijt}$, respectively, where n_{sgt} is the count of students:[32]

$$\left(\overline{A}_{sgt} - \overline{A}_{sgt-1}\right) = \alpha + \lambda_{CFR} \left(\overline{VA}_{sgt} - \overline{VA}_{sgt-1}\right) + \left(\upsilon_{sgt} - \upsilon_{sgt-1}\right) \tag{15}$$

Under the null hypothesis of no forecast bias, we expect $\lambda_{CFR} = 1$, where the average changes in teacher estimates of value-added should correctly forecast average changes in test scores.

Table D2 presents results of the test for a level-only model in Columns (1) to (3) and our matching model in Columns (4) to (6), where each column corresponds to specifications with different levels of controls. With the exception of the level-only model in Column (1),

[32]Following Rothstein (2017) if a teacher is only observed in the omitted year, a zero is assigned for the value-added of these teachers rather than dropping them.

Table D2: Chetty et al. (2014) Quasi-Exp. Test of Forecast Unbiased

	Level-Only Model			Matching Model		
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\lambda}_{CFR}$	1.217	1.007	1.072	1.129	0.922	0.984
95% CI	[1.071, 1.363]	[0.900 , 1.115]	[0.898, 1.245]	[0.983, 1.276]	[0.820, 1.025]	[0.818, 1.150]
School x Year FE	No	Yes	No	No	Yes	No
School x Year FE x Subject	No	No	Yes	No	No	Yes

Note: 22,125 observations. VA in the level-only model consists of the seven teacher-year level effects. VA in the matching model corresponds to Eq. (6). All specifications include controls for grade fixed effects. Regressions are weighted by the number of students in the school-grade-subject-year cell. 95% CI generated with standard errors clustered by school-cohort.

both models indicate unbiased forecasts, where we fail to reject the null hypothesis in each specification.[33]

E Accounting for Classroom Level Shocks

A central concern in the value-added literature is the presence of classroom-level shocks that affect student test scores but are not attributable to the teacher. Several methods exist to account for these shocks. One option, given our use of maximum likelihood estimation, is to incorporate additional assumptions about these shocks directly into the estimation process. However, the more common approach in the literature is to adjust the parameters for these shocks post-estimation, which we will focus on below. Consistent with the literature, we concentrate on classroom shocks that are independent over time.

Our empirical approach addresses both general classroom-level shocks and student-group-

[33]That both models pass the test proposed in Chetty et al. (2014) is not surprising. For example, Delgado (2021) also estimates a matching model with a different dataset and shows that both the level-only model and matching model pass the test. One plausible explanation is that if teachers are assigned students with similar characteristics over time, average value-added will be relatively stable which can be effectively approximated by the level-only model. Thus, the main benefit of the matching model is improved forecasts at the student level, as shown by lower MSE and Log Score in Table 2.

specific shocks in different ways. We assume that the match coefficients are fixed over time, allowing the match effects to be estimated by pooling data across different classrooms. By combining data from multiple years, any varying shocks affecting different groups of students in different years are averaged out as more years are incorporated into the estimation.

As noted by Chetty et al. (2014), assuming any component of value-added is fixed over time can potentially underestimate its impact if it is, in fact, time-varying. Therefore, this more conservative approach to addressing classroom-level shocks that affect match effects is likely to understate our estimates of the importance of matching, as shown in Table 3.

We address general classroom-level shocks similarly to Chetty et al. (2014) by including time-varying teacher-level effects. In this case, the raw estimates of the variance of the teacher-level effects reported in Table 3 may overstate the importance of the level effect because the year effects encompass both the teacher's level effect and the variance of the classroom shock. To estimate the variance of the classroom effect, we follow the strategies outlined by Chetty et al. (2014), such as analyzing auto-covariances by using the one-period covariance of the year effects as a lower bound on the variance of the value-added.

Thus, our population estimates take a conservative approach to addressing classroom-level shocks, which, if anything, may understate our findings on the importance of matching.

Researchers may also be interested in creating estimates of value-added predictors along the lines of Eq. (5) that exclude classroom-level shocks. These estimators can be constructed by using a subset or transformation of the data. Notationally, the complete data is represented by Y_j and Z_j, but the posterior is instead constructed with $\widetilde{Y}_j$ and $\widetilde{Z}_j$, giving $p(\delta_j|\widetilde{Y}_j, \widetilde{Z}_j, \hat{\sigma}^2, \widehat{\Psi})$.

A common approach to remove the influence of general classroom-level shocks is the leave-year-out estimators proposed by Chetty et al. (2014). If classroom shocks are independent, the test score residuals containing these shocks will not be correlated from one year to the next. Thus, removing data from a particular year and using data from other years provides predictions absent of the classroom shocks. In this case, $\widetilde{Y}_j \subset Y_j$ and $\widetilde{Z}_j \subset Z_j$ are subsets of

the original data, and the posterior is constructed using only these subsets.

One drawback of the leave-year-out approach is that it sets aside a significant amount of data, which can affect the estimates of the match components. First, as mentioned, student-group-specific classroom shocks are addressed by averaging over many years. If fewer years are used in the averaging, this will lead to noisier estimates. Second, as shown in Figure 4, the precision of the match components is significantly more sensitive to the amount of available data, so unnecessarily leaving out data will significantly affect the precision of the match components more than that of the level effects.

A possible alternative to the leave-out predictors is to use a partial-out approach instead. The goal of the partial-out approach is to simply remove the source of undesirable variation in the data rather than removing the data altogether. For example, to remove the general classroom effects that may contaminate the teacher year effects, rather than dropping all of the observations for that particular year, the year effects can instead be partialed out of the residuals Y and the covariates Z, allowing the variables that do not contain the classroom shock in that year to contribute to the predictor. Let Z_j^{p} be a subset of the variables in Z to partial out. Then the transformed $\widetilde{Y}_j = Y_j - Z_j^{\mathrm{p}}((Z_j^{\mathrm{p}})'Z_j^{\mathrm{p}})^{-1}Z_j^{\mathrm{p}}Y_j$ and $\widetilde{Z}_j = Z_j - Z_j^{\mathrm{p}}((Z_j^{\mathrm{p}})'Z_j^{\mathrm{p}})^{-1}Z_j^{\mathrm{p}}Z_j$ can be used to construct the posterior distribution, which removes the variation in the data that can be attributed to the variables in Z_j^{p}.

Finally, in certain analysis, for example, studying teachers' impacts on adult outcomes, where the posterior distributions are used as right-hand side variables in a regression, researchers need to be conscientious that the left-hand side variables are not used in the construction of these estimators to avoid a mechanical bias in the regression estimates. For the partial-out estimators, an additional leave-one-out step is likely necessary to remove the sampling error from both sides of the regression. So in this case, the partialling-out removes the classroom-level shock and we only need to leave out one data point to correct for the mechanical correlation of the contemporaneous error with the regressor.

F Matching Model Estimates

Table F1: First Stage Coefficient Estimates

	Math		Reading	
	Coef.	Std. Err.	Coef.	Std. Err.
Math Score$_{t-1}$	f.e.		0.2006***	(0.0020)
Math Score$^2_{t-1}$	f.e.		0.0075***	(0.0008)
Reading Score$_{t-1}$	0.1582***	(0.0018)	f.e.	
Reading Score$^2_{t-1}$	0.0082***	(0.0007)	f.e.	
Female	f.e.		f.e.	
Race Black	f.e.		f.e.	
Race Hispanic	0.0507***	(0.0027)	−0.0188***	(0.0029)
Race Asian	0.1575***	(0.0045)	−0.0051	(0.0047)
Race Other	−0.0171***	(0.0032)	−0.0146***	(0.0035)
FRL	f.e.		f.e.	
LEP	f.e.		f.e.	
(Math Score$_{t-1}$) X (Female)	f.e.		0.0142***	(0.0022)
(Math Score$_{t-1}$) X (Race Black)	f.e.		0.0121***	(0.0028)
(Math Score$_{t-1}$) X (FRL)	f.e.		−0.0193***	(0.0025)
(Math Score$_{t-1}$) X (LEP)	f.e.		−0.0046	(0.0057)
(Reading Score$_{t-1}$) X (Female)	0.0194***	(0.0020)	f.e.	
(Reading Score$_{t-1}$) X (Race Black)	0.0170***	(0.0026)	f.e.	
(Reading Score$_{t-1}$) X (FRL)	−0.0076***	(0.0023)	f.e.	
(Reading Score$_{t-1}$) X (LEP)	0.0038	(0.0054)	f.e.	
(Female) X (Race Black)	f.e.		f.e.	
(Female) X (FRL)	f.e.		f.e.	
(Female) X (LEP)	−0.0009	(0.0080)	0.0079	(0.0088)
(Race Black) X (FRL)	0.0073	(0.0051)	−0.0084	(0.0054)
(Race Black) X (LEP)	0.1119***	(0.0266)	0.1842***	(0.0291)
(FRL) X (LEP)	0.0190	(0.0150)	0.0004	(0.0163)
Teacher Experience	−0.0000	(0.0021)	−0.0006	(0.0023)
Grade 5	0.0222***	(0.0031)	0.0148***	(0.0033)
Ave. Class Math Score$_{t-1}$	−0.0812***	(0.0053)	−0.0321***	(0.0056)
Ave. Class Math Score$^2_{t-1}$	0.0005	(0.0031)	0.0059*	(0.0033)
Ave. Class Reading Score$_{t-1}$	0.0672***	(0.0055)	0.0445***	(0.0059)
Ave. Class Reading Score$^2_{t-1}$	0.0166***	(0.0032)	0.0102***	(0.0034)
Fraction Class Female	0.0233**	(0.0102)	0.0144	(0.0108)
Fraction Class Race Black	−0.0656***	(0.0131)	−0.0329**	(0.0139)
Fraction Class Race Hispanic	−0.0159	(0.0157)	0.0488***	(0.0167)
Fraction Class Race Asian	0.0369	(0.0259)	0.0109	(0.0275)
Fraction Class Race Other	−0.0458**	(0.0200)	−0.0098	(0.0212)

(*Continued on next page*)

Note: This table reports the coefficients and standard errors for the covariates/background characteristics included in the first stage separately for math and reading. FRL denotes Free or Reduced-Price Lunch. LEP denotes Limited English Proficiency. *** denotes significance at the 1%, ** at the 5%, and * at the 10% levels. Covariates indicated with 'f.e.' correspond to variables that are interacted with teacher fix effects.

Table F1: First Stage Coefficient Estimates

	Math		Reading	
	Coef.	Std. Err.	Coef.	Std. Err.
Fraction Class FRL	-0.0484^{***}	(0.0095)	-0.0229^{**}	(0.0101)
Fraction Class LEP	0.0087	(0.0186)	-0.0430^{**}	(0.0198)
Class Size	-0.0088^{***}	(0.0015)	-0.0057^{***}	(0.0016)
Class Size^2	0.0001^{*}	(0.0000)	0.0001^{*}	(0.0000)
Ave. School Math Score$_{t-1}$	-0.1991^{***}	(0.0110)	-0.0776^{***}	(0.0117)
Ave. School Math Score$^2_{t-1}$	0.0114	(0.0081)	-0.0035	(0.0086)
Ave. School Reading Score$_{t-1}$	0.0305^{**}	(0.0132)	-0.0793^{***}	(0.0140)
Ave. School Reading Score$^2_{t-1}$	-0.0478^{***}	(0.0084)	-0.0334^{***}	(0.0089)
Fraction School Female	0.0469^{*}	(0.0247)	-0.0011	(0.0263)
Fraction School Race Black	-0.0890^{***}	(0.0248)	-0.0259	(0.0263)
Fraction School Race Hispanic	0.0105	(0.0361)	-0.0746^{*}	(0.0384)
Fraction School Race Asian	0.1855^{***}	(0.0600)	0.0464	(0.0637)
Fraction School Race Other	0.0639	(0.0500)	0.0023	(0.0531)
Fraction School FRL	-0.2006^{***}	(0.0223)	-0.2279^{***}	(0.0237)
Fraction School LEP	0.0796^{*}	(0.0441)	0.1306^{***}	(0.0469)
School Size	0.0021	(0.0093)	-0.0116	(0.0099)
School Size^2	-0.0029	(0.0018)	0.0006	(0.0019)
Ave. District Math Score$_{t-1}$	0.0841^{***}	(0.0227)	0.0382	(0.0241)
Ave. District Math Score$^2_{t-1}$	-0.0389^{*}	(0.0236)	0.0622^{**}	(0.0251)
Ave. District Reading Score$_{t-1}$	0.0446	(0.0302)	-0.0381	(0.0321)
Ave. District Reading Score$^2_{t-1}$	0.1645^{***}	(0.0238)	0.0201	(0.0253)
Fraction District Female	-0.4638^{***}	(0.0847)	-0.1408	(0.0900)
Fraction District Race Black	-0.0160	(0.0479)	-0.0564	(0.0508)
Fraction District Race Hispanic	0.1226	(0.0819)	-0.1452^{*}	(0.0868)
Fraction District Race Asian	-0.9082^{***}	(0.1971)	-0.0942	(0.2092)
Fraction District Race Other	-0.0495	(0.1057)	-0.0461	(0.1111)
Fraction District FRL	0.0638	(0.0446)	0.0973^{**}	(0.0474)
Fraction District LEP	-0.3392^{***}	(0.0982)	-0.0608	(0.1042)
District Size	-0.0003	(0.0023)	0.0039	(0.0024)
District Size^2	0.0003^{***}	(0.0001)	-0.0000	(0.0001)
Year 2009	-0.0166^{***}	(0.0033)	-0.0037	(0.0035)
Year 2010	0.0093^{*}	(0.0055)	0.0141^{**}	(0.0058)
Year 2011	0.0059	(0.0075)	0.0114	(0.0080)
Year 2012	0.0096	(0.0097)	0.0205^{**}	(0.0104)
Year 2013	0.0114	(0.0101)	0.0333^{***}	(0.0108)
Year 2014	0.0340^{***}	(0.0123)	0.0524^{***}	(0.0132)

Note: This table reports the coefficients and standard errors for the covariates/background characteristics included in the first stage separately for math and reading. FRL denotes Free or Reduced-Price Lunch. LEP denotes Limited English Proficiency. *** denotes significance at the 1%, ** at the 5%, and * at the 10% levels. Covariates indicated with 'f.e.' correspond to variables that are interacted with teacher fix effects.

Table F2: Population Distribution of Value-Added Coefficients: Math

	Component 1		Component 2		Population
	Mean	Std. Dev.	Mean	Std. Dev.	Std. Dev.
	(1)	(2)	(3)	(4)	(5)
Year 2008 (δ_1)	-0.0823	0.2343	0.1584	0.2099	0.2534
	(0.0091)	(0.0060)	(0.0144)	(0.0086)	(0.0032)
Year 2009 (δ_2)	-0.0631	0.2159	0.1215	0.2435	0.2421
	(0.0078)	(0.0056)	(0.0149)	(0.0082)	(0.0031)
Year 2010 (δ_3)	-0.0601	0.2156	0.1157	0.2454	0.2411
	(0.0075)	(0.0056)	(0.0141)	(0.0082)	(0.0032)
Year 2011 (δ_4)	-0.0732	0.2151	0.1408	0.2155	0.2380
	(0.0080)	(0.0057)	(0.0145)	(0.0086)	(0.0033)
Year 2012 (δ_5)	-0.0736	0.2221	0.1418	0.2242	0.2451
	(0.0083)	(0.0058)	(0.0155)	(0.0089)	(0.0033)
Year 2013 (δ_6)	-0.0785	0.2380	0.1512	0.2261	0.2581
	(0.0091)	(0.0064)	(0.0160)	(0.0096)	(0.0035)
Year 2014 (δ_7)	-0.0920	0.2319	0.1772	0.2246	0.2626
	(0.0097)	(0.0066)	(0.0170)	(0.0100)	(0.0035)
A_{t-1} (δ_8)	0.0042	0.0515	-0.0080	0.0720	0.0596
	(0.0030)	(0.0045)	(0.0055)	(0.0058)	(0.0025)
A_{t-1}^2 (δ_9)	0.0126	0.0310	-0.0243	0.0397	0.0384
	(0.0016)	(0.0018)	(0.0028)	(0.0022)	(0.0009)
FEM (δ_{10})	0.0052	0.0461	-0.0100	0.0650	0.0538
	(0.0041)	(0.0091)	(0.0072)	(0.0099)	(0.0046)
BL (δ_{11})	0.0165	0.0418	-0.0317	0.0895	0.0664
	(0.0048)	(0.0143)	(0.0089)	(0.0111)	(0.0058)
FRL (δ_{12})	0.0069	0.0630	-0.0134	0.0712	0.0666
	(0.0044)	(0.0076)	(0.0079)	(0.0109)	(0.0045)
LEP (δ_{13})	0.0039	0.0679	-0.0075	0.0701	0.0689
	(0.0070)	(0.0137)	(0.0126)	(0.0225)	(0.0086)
(A_{t-1} × FEM) (δ_{14})	-0.0030	0.0301	0.0058	0.0459	0.0365
	(0.0029)	(0.0077)	(0.0052)	(0.0080)	(0.0039)
(A_{t-1} × BL) (δ_{15})	-0.0052	0.0392	0.0100	0.0697	0.0522
	(0.0037)	(0.0093)	(0.0069)	(0.0096)	(0.0046)
(A_{t-1} × FRL) (δ_{16})	-0.0010	0.0244	0.0019	0.0584	0.0395
	(0.0033)	(0.0116)	(0.0060)	(0.0085)	(0.0045)
(A_{t-1} × LEP) (δ_{17})	0.0040	0.0905	-0.0076	0.0918	0.0911
	(0.0068)	(0.0111)	(0.0130)	(0.0204)	(0.0075)
(FEM × BL) (δ_{18})	-0.0046	0.0397	0.0089	0.0887	0.0614
	(0.0060)	(0.0241)	(0.0107)	(0.0181)	(0.0099)
(FEM × FRL) (δ_{19})	-0.0002	0.0643	0.0005	0.0515	0.0603
	(0.0057)	(0.0131)	(0.0101)	(0.0262)	(0.0085)
Component Share	0.6581		0.3419		

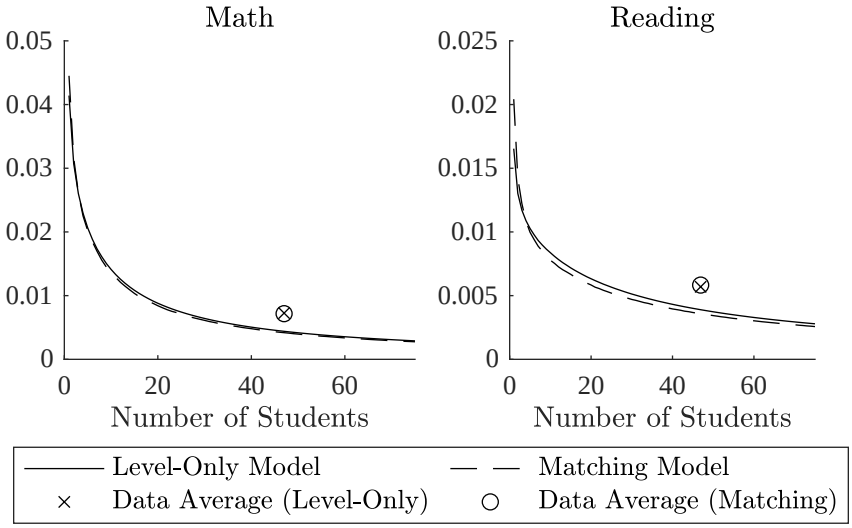
Note: This table presents the mean and standard deviation of the math value-added coefficients for each component of the mixture, as well as for the overall population. The population means are normalized to zero, with standard errors provided in parentheses.

Table F3: Population Distribution of Value-Added Coefficients: Reading

	Component 1		Component 2		Population
	Mean	Std. Dev.	Mean	Std. Dev.	Std. Dev.
	(1)	(2)	(3)	(4)	(5)
Year 2008 (δ_1)	0.0022	0.1538	-0.0590	0.2974	0.1616
	(0.0030)	(0.0033)	(0.0610)	(0.0550)	(0.0041)
Year 2009 (δ_2)	0.0013	0.1387	-0.0344	0.3038	0.1480
	(0.0028)	(0.0033)	(0.0517)	(0.0478)	(0.0040)
Year 2010 (δ_3)	0.0033	0.1336	-0.0877	0.3285	0.1462
	(0.0028)	(0.0034)	(0.0596)	(0.0467)	(0.0043)
Year 2011 (δ_4)	0.0011	0.1375	-0.0303	0.2863	0.1457
	(0.0029)	(0.0035)	(0.0611)	(0.0498)	(0.0042)
Year 2012 (δ_5)	0.0014	0.1400	-0.0369	0.2678	0.1467
	(0.0030)	(0.0034)	(0.0677)	(0.0615)	(0.0046)
Year 2013 (δ_6)	0.0037	0.1406	-0.1003	0.2299	0.1460
	(0.0031)	(0.0038)	(0.0681)	(0.0685)	(0.0044)
Year 2014 (δ_7)	0.0011	0.1452	-0.0306	0.2718	0.1517
	(0.0032)	(0.0039)	(0.0728)	(0.0833)	(0.0056)
A_{t-1} (δ_8)	0.0025	0.0604	-0.0664	0.1881	0.0704
	(0.0016)	(0.0027)	(0.0349)	(0.0269)	(0.0032)
A_{t-1}^2 (δ_9)	-0.0027	0.0293	0.0736	0.0649	0.0344
	(0.0008)	(0.0012)	(0.0149)	(0.0106)	(0.0014)
FEM (δ_{10})	-0.0004	0.0325	0.0096	0.1826	0.0471
	(0.0022)	(0.0085)	(0.0434)	(0.0546)	(0.0081)
BL (δ_{11})	0.0025	0.0432	-0.0662	0.1612	0.0538
	(0.0026)	(0.0094)	(0.0479)	(0.0726)	(0.0093)
FRL (δ_{12})	0.0019	0.0514	-0.0499	0.2197	0.0661
	(0.0023)	(0.0063)	(0.0443)	(0.0520)	(0.0066)
LEP (δ_{13})	0.0002	0.0418	-0.0055	0.2772	0.0667
	(0.0045)	(0.0208)	(0.0856)	(0.0961)	(0.0170)
(A_{t-1} × FEM) (δ_{14})	-0.0011	0.0406	0.0285	0.1336	0.0475
	(0.0016)	(0.0040)	(0.0290)	(0.0307)	(0.0041)
(A_{t-1} × BL) (δ_{15})	0.0000	0.0541	-0.0011	0.1536	0.0605
	(0.0021)	(0.0052)	(0.0392)	(0.0531)	(0.0058)
(A_{t-1} × FRL) (δ_{16})	0.0002	0.0573	-0.0042	0.1330	0.0616
	(0.0019)	(0.0040)	(0.0380)	(0.0433)	(0.0044)
(A_{t-1} × LEP) (δ_{17})	-0.0005	0.0799	0.0137	0.3296	0.1003
	(0.0041)	(0.0104)	(0.0859)	(0.0828)	(0.0117)
(FEM × BL) (δ_{18})	-0.0013	0.0302	0.0340	0.1378	0.0400
	(0.0033)	(0.0229)	(0.0623)	(0.1165)	(0.0192)
(FEM × FRL) (δ_{19})	-0.0007	0.0293	0.0193	0.2556	0.0564
	(0.0030)	(0.0194)	(0.0610)	(0.0700)	(0.0130)
Component Share	0.9641		0.0359		

Note: This table presents the mean and standard deviation of the reading value-added coefficients for each component of the mixture, as well as for the overall population. The population means are normalized to zero, with standard errors provided in parentheses.

Figure G1: Mean Variance of Predictive Distribution of Value-Added By Number of Observations



G Additional Results on the Precision of Value-added Predictors

G.1 Comparison of Precision of a Level-Only Model to the Matching Model

Figure G1 compares the precision of the level-only value-added model with the matching model for a given number of observed students. To make the comparison between the two models, we sample n students for 1,000 teachers for 0 to 75 students. For each teacher, we calculate the precision of their value-added for the average student they are assigned, i.e., $\text{Var}(VA_{ij}|\overline{z}_j)$. The figure then plots the mean of the average precision across the 1,000 teachers for $n \in [1, 75]$.

The main takeaway from Figure G1 is that for both subjects, the precision of the matching model for the average student is nearly identical to that of the level-only model. Thus, although the matching model is more complex there does not appear to be any loss of precision. The figure also plots the average value of the precision in the data for the two models. Teachers in our data have on average 47 students. The precision of the data

average is slightly higher than the simulations because in the simulations we randomly sample students across years, whereas in the data, students are organized into classes that are concentrated by year.

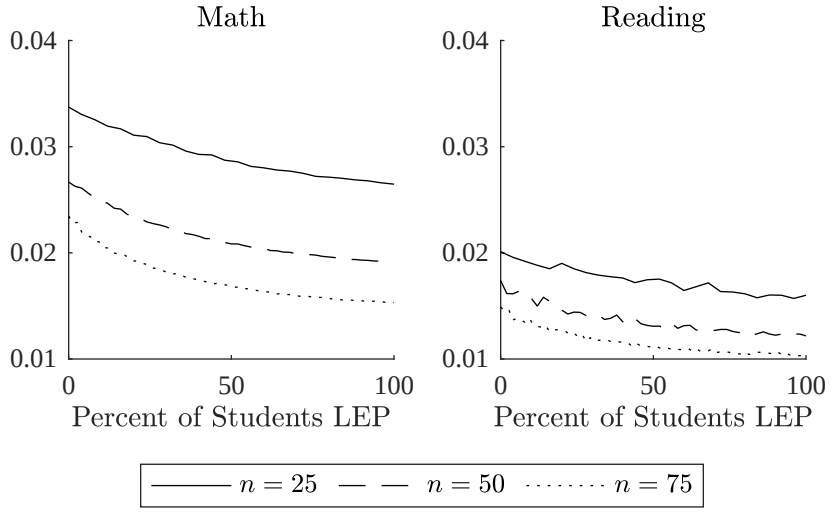
G.2 Precision of Matching Model For Sparsely Distributed Student Characteristics

Figure G1 focuses on the precision of the value-added estimates for the average student assigned to a teacher. However, in the matching model, a teacher's value-added is student-specific. Thus, we are not only concerned with the precision on average but also with the precision for a given student. In particular, some teachers may have very little data pertaining to their value-added for certain student groups. Even though it is possible to make predictions of the teacher's value-added for student types when there is limited data, it is important to understand the level of precision of these predictions.

Here we focus on Limited English Proficiency (LEP) students as an example and examine how the number and characteristics of the students assigned to a teacher affect the precision of the predicted value-added that a teacher has specifically for LEP students. LEP students comprise a small minority of the student population in North Carolina, and it is not unusual to see teachers with very few LEP students. However, as shown in Figure 2, LEP students are one of the groups most impacted by teacher assignment, so it is useful to be able to assess the precision of a teacher's value-added predictor when there is very little data on the teacher's experience with LEP students.

In our model, the teacher's value-added not only depends on whether the student is identified as LEP but also varies across LEP students with different prior achievement levels, as the lag score is interacted with LEP. Thus, in this analysis we will focus on the teachers average precision of their value-added for LEP students, which takes the form $\text{Var}(VA_{ij}|LEP_i = 1) = \int \text{Var}(VA_{ij}|z)g(z|LEP_i = 1)dz$, where $g(z|LEP_i = 1)$ is the conditional distribution of the student characteristics z given that the student has $LEP = 1$.

Figure G2: Mean Variance of Predictive Distribution of Value-Added For LEP Students



There are two channels that allow us to make predictions about a teacher's value-added for LEP students, even if they have only taught few LEP students in the past. The first is a direct channel. Limited English Proficiency is just one dimension of many that describes LEP students. LEP students also differ in term of prior academic achievement, gender, race, and FRL-status. Thus, any informative data available for the teacher in these dimensions can be applied to make predictions for LEP students. The second is an indirect channel related to the teacher's comparative advantage through the distribution of δ. For example, if teachers with a comparative advantage for FRL students tend to also have a comparative advantage for LEP students, this will be reflected in the correlation of the distribution of δ.

First, the variance of the prior of value-added for LEP students is 0.0665 for math and 0.0266. Figure G2 shows the average precision of value-added for LEP students for teachers with 25, 50 and 75 students. We also, vary the percent of LEP students from 0 to 100. For example, a teacher with 50 students where none of them are LEP is able to reduce the amount of uncertainty of their value-added for LEP students by 59.8% in math (100*(1 - 0.0267/0.0665)) and 40% in reading (100*(1 - 0.0161/0.0266)). In fact this reduction is even larger than the improvement in precision for a teacher with 25 students who are all LEP.

Thus this figure demonstrates that it is not strictly necessary for a teacher to be assigned

many LEP students to improve the precision of their value-added for these students, but the total number of students assigned also plays an important role, even if they are non-LEP.

H Certainty Equivalent Value-Added to Rank Teachers

We propose a transparent approach to rank teachers that avoids the idiosyncrasies that may arise from realized classroom assignments. One method to assess teachers and their value-added across a portfolio of possible classrooms is to view the administrator's problem from an expected utility framework. Under this approach, because it may be uncertain to which classroom a teacher will be assigned, the teacher's contribution to student test scores is indeterminate.[34] The administrator may consider circumventing this uncertainty and replacing this teacher with a candidate with a value-added that is invariant to class assignments.[35] We are interested in calculating the precise constant value-added for the candidate that would make the administrator indifferent between keeping the teacher with the portfolio of value-added and replacing her with the risk-free candidate. This value depends on the administrator's objectives and represents the teacher's certainty equivalent value-added (CEVA). As Chetty et al. (2014) and Hanushek (2009) suggest threshold rules based on traditional level-only value-added rankings, similarly, administrators instead could make personnel decisions based on certainty equivalent thresholds, for example dismissing the teachers who have a certainty equivalent below the 5th percentile.

We assume that the administrator has exponential utility, so the calculation of the certainty equivalent is based on two components: the set of classrooms to which a teacher could be assigned and the administrator's risk preference, ρ.[36] Specifically, let S be the number of

[34]Uncertainty naturally arises from year-to-year turnover along with other forces that impact labor force composition, personnel assignment decisions, and entry and exit of cohorts of students.

[35]Such a teacher would offer guaranteed identical returns in any classroom.

[36]Exponential utility with risk preference ρ is given by $u(x) = -\exp(-\rho x)$.

possible classroom assignments within a district or state, and $\overline{VA}_{js} = \mathrm{E}(VA_{js}) = \overline{z}'_s \hat{\delta}_j^{BLUP}$ is teacher j's expected value-added upon assignment to class s. The probability that teacher j is assigned to class s is denoted p_{js} with $\sum_{s=1}^{S} p_{js} = 1$. These probabilities are determined by a number of factors which include the ability to move teachers across schools or grades, the distribution of the value-added for the other teachers in the school or district, and the general assignment practices of the administrator, among other things.[37] Because different administrators may have different tolerances for risk, the certainty equivalent is also a function of the administrator's risk preference ρ, where $\rho = 0$ is risk neutral, $\rho > 0$ is risk-averse, and $\rho < 0$ is risk-seeking.[38] Together the certainty equivalent value-added for teacher j is determined by

$$CEVA_j = \begin{cases} \ln\left[\left(\sum_{s=1}^{S} p_{js} \exp\left(\overline{VA}_{js}\right)^{-\rho}\right)^{(1/-\rho)}\right] & \text{if } \rho \neq 0 \\ \sum_{s=1}^{S} p_{js}\left(\overline{VA}_{js}\right) & \text{if } \rho = 0 \end{cases} \tag{16}$$

To offer some insight into the workings of Eq. (16), Table H1 provides a simple two-classroom example. In this example, if the teacher is assigned to the first classroom, her average value-added is -0.05. Alternatively upon matching with the second classroom, the teacher's average value-added is 0.20. Table H1 illustrates the CEVA for this teacher under a number of situations. The first row of the table shows the CEVA when the teacher has an equal probability of being assigned to each class and the administrator is risk neutral. In this case, the CEVA is just the average of the value-added across the two classes, 0.075. The second and third rows show cases where the teacher is assigned with certainty to one class. In these cases, the CEVA approaches the value-added associated with the class assignment, and

[37] These probabilities could reasonably be assumed to be exogenous for the administrator if the hiring/firing policies and teacher allocation decisions are implemented by different levels of the school administration system (e.g., district vs. school level).

[38] We show later that risk-averse administrators are more aggressive in terminating teachers. Risk is defined as the possibility that the current teacher assignment to another class yields a poor match, which are outcomes the risk-averse administrator prefers to avoid.

Table H1: Two Classroom Example of Certainty Equivalent Rank Calculation

Teacher value-added...
... if assigned to first classroom $\overline{VA}_{j1} = -0.05$
... if assigned to second classroom $\overline{VA}_{j2} = 0.20$

p_1	p_2	ρ	$CEVA$
1/2	1/2	0	0.075
0	1	Any Value	0.20
1	0	Any Value	-0.05
Non-Zero	Non-zero	∞	-0.05
Non-Zero	Non-zero	$-\infty$	0.20

Note: Examples of how classroom allocation could impact certainty equivalent rank calculation.

the risk preference of the administrator does not matter. The last two rows of Table H1 show how CEVA is calculated with different risk tolerances. In the case of infinite risk-aversion, $\rho = \infty$, the administrator will rank teachers based on the teacher's lowest possible match that occurs with non-zero probability, while in the infinitely risk-seeking case, $\rho = -\infty$, the rank of the teacher will be based on the highest match that occurs with positive probability. A very risk-averse administrator would be willing to replace this teacher with one whose invariant value-added was slightly higher than the worse match of -0.05. On the other hand, a risk-seeking administrator would only be satisfied with a replacement teacher whose invariant-value-added approached the higher match value of 0.20.

The certainty equivalent approach generalizes a number of intuitive ways that an administrator may prefer to rank teachers. For example, if teachers are ranked based on their average value-added across all available classrooms in the district, this corresponds to giving equal weight to each classroom (i.e. $p_{js} = 1/S$ for all s) and risk neutrality. In addition, the certainty equivalent approach nests ranking based on a level-only value-added model, which assumes that $p_{js^o} = 1$ for teacher j, where s^o is the class that they taught in the estimation sample, that is with certainty teachers will teach identical students in the future as they

taught in the past.

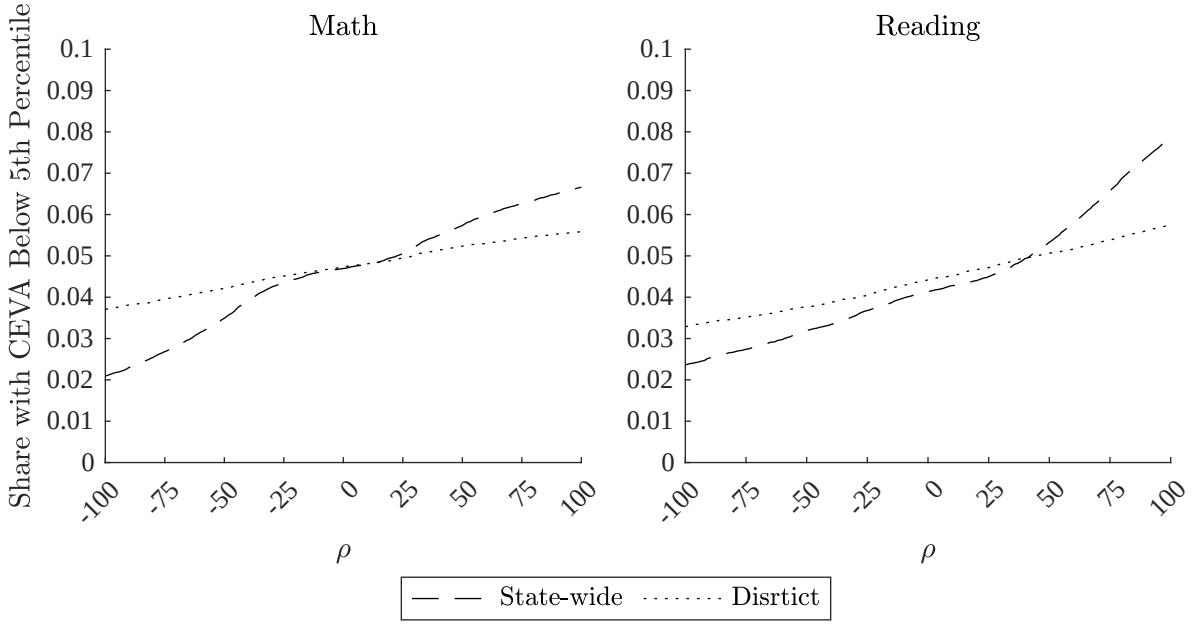
To conclude this section, we use the same population of teachers in Table 5 and compare their certainty equivalent value added to the 5% firing thresholds, -0.1565 for reading and -0.3014 for math. For each teacher, we construct the CEVA under two potential classroom assignment schemes, uniform assignment to classrooms at the district level and the state level (i.e., $p_{js} = 1/S$ for all s, where S corresponds to the total number of classrooms at the district or state). We view these thresholds as minimum teacher requirements, and we are interested in calculating the share of teachers with a certainty equivalent value-added below these thresholds if they were randomly assigned a classroom at the district or state level.

Figure H1 displays the fraction of teachers with CEVA below the value-added thresholds for risk preference parameters in the range $[-100, 100]$. This figure shows the administrators degree of risk preference can lead to considerably different firing rates. For example, a highly risk-averse administrator ($\rho = 100$), seeking to mitigate the possibility of a low value-added outcome would dismiss 5.7% of teachers for reading when CEVA is calculated at the district level. On the other-hand, a risk-seeking administrator ($\rho = -100$), who places more utility on high value-added outcomes would only dismiss 3.3% of teachers. The gap in firing outcomes are even greater using the state-level assignment probabilities for reading because there is more classroom variation, increasing the probability of both high and low matches. The risk-averse administrator would respond to the increased probability of low matches by firing more teachers, approximately 7.9%, while the risk-seeking administrator would respond to the increased probability of high matches by firing less than 2.4% of the teachers.

It is notable that the plot for math in Figure H1 is generally flatter than the plot for reading. This is due to the fact that the variance of the level effect as shown in Table 3 is much larger for math than reading. The level effect's role in the certainty equivalent calculation is an additive constant that is independent of the risk parameter.[39] Thus, because the variance

[39]Given that $\overline{VA}_{js} = \delta_{j0} + \sum_{k=1}^{K} \overline{z}_{sk}\delta_{jk}$, where $\overline{z}_{sk}$ is class s's average of attribute k, Equation (16) can

Figure H1: Share of Teachers with CEVA Below -0.1565 for Reading and -0.3014 for Math value-added Threshold



Note: The thresholds -0.1565 and -0.3014, are the 5th percentiles in the value-added distribution of the traditional level-only value-added model for reading and math respectively, which Chetty et al. (2014) and Hanushek (2009) have proposed as being the basis for dismissing teachers.

of the level effect is so large in math, the level effect plays a much greater role and brings more stability in CEVA rankings even as ρ changes. For reading, the level effect plays a smaller role because it has a smaller variance. In this case, the CEVA ranking primarily reflects the match components, which is impacted by the risk parameter, generating big differences in the rankings for different values of ρ.

This analysis demonstrates that because student classroom compositions vary, a given teacher's value-added will be different in each classroom due to the match effects, which makes measuring teacher effectiveness difficult. In an environment where teachers transfer schools and districts, leading to unpredictable composition changes in their classes across

also be written as:

$$CEVA_j = \begin{cases} \delta_{j0} + \ln\left[\left(\sum_{s=1}^{S} p_{js} \exp\left(\sum_{k=1}^{K} \overline{z}_{sk}\delta_{jk}\right)^{-\rho}\right)^{(1/-\rho)}\right] & \text{if } \rho \neq 0 \\ \delta_{j0} + \sum_{s=1}^{S} p_{js}\left(\sum_{k=1}^{K} \overline{z}_{sk}\delta_{jk}\right) & \text{if } \rho = 0 \end{cases}$$

time, even a very simple approach that measures a teacher's effectiveness as their average value-added across all classrooms (i.e., CEVA with $p_{js} = 1/S$ for all s, and risk neutrality) could provide a more robust measure to compare teachers and may lead to better personnel outcomes than the current approaches suggested in the value-added literature.[40]

[40]Over 30% of teachers in North Carolina express interest in searching for a new position, and between 10 to 20% of teachers leave their current position each year, whether through transfers or quitting the profession. Finally, if administrators can predict teacher movement, the probability weights may be set with more purpose than merely equal probability assignment. Moreover, administrators who are confident that they have policies and personnel in place at the school level to match teachers to their best classes may dial up their risk preference and raise their CEVA, ultimately terminating fewer teachers by placing more teachers in environments where they can be successful.