NBER WORKING PAPER SERIES

A NEW METHOD FOR ESTIMATING TEACHER VALUE-ADDED

Michael Gilraine
Jiaying Gu
Robert McMillan

Working Paper 27094
http://www.nber.org/papers/w27094

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2020

We would like to thank Roger Koenker for helpful discussions, and Joe Altonji, Raj Chetty, Michael Dinerstein, John Friedman, Chris Taber, Sergio Urzua and seminar participants at the NBER Summer Institute, Yale University, the University of Maryland, Western, the Banff International Research Station workshop, and the Jacobs Center CCWD workshop for additional comments. Guan Yi Lin and Hammad Shaikh provided excellent research assistance. Financial support from SSHRC and the University of Toronto Mississauga is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A New Method for Estimating Teacher Value-Added
Michael Gilraine, Jiaying Gu, and Robert McMillan
NBER Working Paper No. 27094
May 2020
JEL No. C14,H75,I21,J24,J45

ABSTRACT

This paper proposes a new methodology for estimating teacher value-added. Rather than imposing a normality assumption on unobserved teacher quality (as in the standard empirical Bayes approach), our nonparametric estimator permits the underlying distribution to be estimated directly and in a computationally feasible way. The resulting estimates fit the unobserved distribution very well regardless of the form it takes, as we show in Monte Carlo simulations. Implementing the nonparametric approach in practice using two separate large-scale administrative data sets, we find the estimated teacher value-added distributions depart from normality and differ from each other. To draw out the policy implications of our method, we first consider a widely-discussed policy to release teachers at the bottom of the value-added distribution, comparing predicted test score gains under our nonparametric approach with those using parametric empirical Bayes. Here the parametric method predicts similar policy gains in one data set while overestimating those in the other by a substantial margin. We also show the predicted gains from teacher retention policies can be underestimated significantly based on the parametric method. In general, the results highlight the benefit of our nonparametric empirical Bayes approach, given that the unobserved distribution of value-added is likely to be context-specific.

Michael Gilraine
New York University
Department of Economics
19 West 4th Street
New York, NY 10012
mike.gilraine@nyu.edu

Jiaying Gu
Department of Economics
University of Toronto
150 St. George Street
Toronto, ON M5S 3G7
Canada
jiaying.gu@utoronto.ca

Robert McMillan
University of Toronto
Department of Economics
150 St. George Street
Toronto, ON M5S 3G7
CANADA
and NBER
mcmillan@chass.utoronto.ca

1 Introduction

Measuring the impact of teachers on student achievement has been a longstanding preoccupation in applied research – naturally so, given the vital role that teachers play in education production. As observable characteristics tend to do a poor job when predicting teacher performance,[1] researchers have proposed influential fixed effects methods as a means to capture a teacher's overall quality, taking advantage of large-scale matched student-teacher data sets that are increasingly accessible – see pioneering studies by Rockoff (2004) and Rivkin, Hanushek, and Kain (2005). In turn, fixed effects methods have prompted the development of teacher value-added ('VA') estimators for measuring the impact of teachers that are both transparent and easy to implement.[2] Given the appeal of such estimators, teacher VA estimates now feature ever more widely in the policy sphere, particularly in consequential teacher retention, promotion and pay decisions. Indeed, by the end of 2017, fully thirty nine states required VA measures to be incorporated into teacher evaluation scores (as one indicator of this phenomenon).

The use of VA methods in high-stakes decision making raises important challenges. Not least, such methods need to be able to recover teacher quality on the basis of relatively few teacher-year observations, particularly so for teachers new to the profession. The standard approach to this issue involves using empirical Bayes methods to reduce measurement error in VA estimates, 'shrinking' less reliable estimates back toward the mean (Kane and Staiger, 2008; Kane et al., 2008; Jacob and Lefgren, 2008; Harris and Sass, 2014; Chetty et al., 2014a,b). In order to apply these methods, papers estimating teacher VA have typically used the parametric empirical Bayes ('PEB') estimator, first proposed by James and Stein (1961). This is attractive given its analytic convenience and also because it is the feasible version of the optimal Bayes rule for estimating teacher quality when unobserved quality is normally distributed.[3]

[1] For instance, Kane, Rockoff, and Staiger (2008) show that among teachers with identical experience and certification status, there are large and persistent differences in teacher effectiveness.

[2] See Koedel, Mihaly, and Rockoff (2015) for a recent review.

[3] To be precise, it is the feasible version of the parametric Bayes estimator – the optimal Bayes rule under

In practice, unobserved teacher quality may not follow a normal distribution. Given this possibility, we do not have a clear sense of how the resulting VA estimates might be affected by departures from normality, nor of the implications that such departures could have for policies based on VA estimates. The analysis in this paper seeks to shed light on both these relevant issues.

The central contribution of our paper is to set out a feasible new methodology for estimating teacher VA – one that does not impose any parametric assumptions on the unobserved heterogeneity in teacher quality. Following a standard setup in which residualized test scores equal underlying teacher quality (the heterogeneous teacher VA of interest) plus noise, we show first that the teacher quality distribution can be identified nonparametrically – see Theorem 1 below.[4] Next, we derive the nonparametric Bayes estimator for teacher VA (see Theorem 2), drawing on a path-breaking 1956 paper in statistics by Herbert Robbins.[5] This nonparametric estimator is optimal in the sense of minimizing the mean squared error of individual teacher quality estimates *regardless* of the true underlying distribution of unobserved teacher quality.

In terms of shrinkage, both the parametric and nonparametric Bayes estimators can be written as functions of teacher fixed effects. Whereas the standard parametric Bayes estimator shrinks the teacher fixed effect linearly, our nonparametric Bayes estimator features a non-linear shrinkage rule, allowing the amount of shrinkage applied to each teacher fixed effect to be non-monotonic. This estimator is infeasible, however, given it involves the unknown teacher VA distribution, denoted F. We obtain a feasible version by applying a two-step approach, described in Section 3.[6] Doing so yields the nonparametric empirical

quadratic error loss in the case where unobserved teacher quality follows a normal distribution.

[4]In Appendix A, we provide a general deconvolution proof of nonparametric identification in the case of teacher VA. In the main analysis, for tractability, we make the assumption that the noise component in the residualized test score model is independent of underlying teacher quality and has a known distribution, assumed to be normal with a common variance. To assume that unobserved teacher quality in this formulation also follows a normal distribution involves an over-parameterization.

[5]In the words of Efron (2003), "There seems to be a good chance that Robbins was 50 years ahead of his time and that a statistical theory of the 1950s will shine in the 21st century." (See page 377.)

[6]In essence, we first estimate the teacher VA distribution F, using nonparametric maximum likelihood, to get $\hat{F}$, and second, plug this into the first equation in Theorem 2 (the theorem defining the nonparametric Bayes estimator).

Bayes ('NPEB') estimator used in the main analysis.

Implementing the Robbins nonparametric Bayes approach in large-scale empirical applications has only recently become viable, following important computational advances by Koenker and Mizera (2014).[7] We leverage those advances in the current study, showing first that the NPEB approach performs very well in Monte Carlo simulations in an environment that mimics typical administrative data sets closely, being responsive to the true underlying VA distribution. This contrasts with the standard PEB approach, which becomes less reliable when departures from normality are more pronounced, as the simulations demonstrate.

Next we apply the methodology using observational data, estimating teacher VA in two separate large-scale administrative data sets. One covers the entire state of North Carolina and the other, the Los Angeles Unified School District (LAUSD) – the second largest school district in the United States. The estimated teacher VA distributions differ both from the normal distribution assumed in the prior literature and from each other. In North Carolina, the estimated distribution has a relatively similar shape to the normal, although with fatter tails;[8] in the LAUSD, our estimated teacher distribution is skewed, with a much thinner left than right tail. In each instance, the deviations from normality are statistically significant, as indicated by a new diagnostic test we develop (see Appendix E).

Given the deviations from normality in both settings, we then evaluate the policy relevance of our methodology. To the extent that the normality assumption is misplaced, our approach will improve the prediction of the total gains of a particular policy (relative to the standard PEB method); such gains could then be weighed by decision-makers against total costs in trying to determine optimal policy. First we consider the total gains of a widely discussed proposal to release teachers. The literature has focused on teachers in the bottom five percent of the estimated teacher value-added distribution, as outlined by Hanushek (2009, 2011) and evaluated in Chetty et al. (2014b).[9] Accordingly, we compare the predicted test

[7]Other applications in economics include analyses of earnings dynamics – see Gu and Koenker (2017b), for instance.

[8]This finding aligns with Goldhaber and Startz (2017) who find that the distribution of teachers in North Carolina is not Gaussian, but the differences from the normal distribution tend to be small.

[9]Because empirical Bayes seeks to minimize prediction error, we focus on the out-of-sample predictions of

score gains of students using our method with an approach that imposes normality on the underlying teacher quality distribution, supposing the bottom five percent of teachers (based on estimated VA after three years of observation) are released and replaced by teachers of average quality.

In North Carolina, we find only minor differences between the two approaches: the PEB method overstates test score gains of the policy by around five percent relative to our methodology. In contrast, the skewness of the distribution of teacher value-added in the LAUSD leads to large differences in the estimated policy benefit, with PEB overstating test score gains of the policy by over a quarter.

More broadly, we are able to simulate the aggregate test score effects of policies that release *any* given percentage of teachers from the bottom of the VA distributions under the two approaches. We also include classroom fixed effects, and repeat the estimation and the policy analyses on that basis. Doing so reduces the variance in the VA estimates, as one would expect, and lowers the extent to which PEB overstates the test score gains of teacher release policies. Still, using LAUSD data, PEB overstates the gains by 16 percent under the benchmark '5 percent cutoff' policy. When we consider 'teacher retention' policies, in contrast, we show that PEB can underestimate policy gains, the underestimation becoming more severe in North Carolina when including classroom fixed effects.

At a general level, our analysis underscores the plausible notion that the true underlying distribution of teachers is likely to be context-specific, and so estimated policy gains when invoking normality may well differ substantially from the true policy gains in some settings. Here, our data-driven methodology offers policymakers a flexible means to understand the benefits of implementing the same reform in different environments. Further, given the computational feasibility of our approach, analytical convenience need no longer weigh on the side of assuming normality. This opens up new possibilities for empirical research, not

policies that are relevant to understanding how *many* teachers to release. Given empirical Bayes does not seek to minimize teacher ranking errors, it is less useful for studying teacher rankings – that is, deciding *whom* to release. Indeed, we will find that the choice of estimators – whether fixed effect, parametric, or nonparametric empirical Bayes – has little appreciable impact on teacher rankings in either our simulations (see Tables F.1(a)-F.1(c)) or empirical applications (see Section 6.1).

least as the nonparametric empirical Bayes methodology is applicable in a variety of other contexts where parametric empirical Bayes methods have already been used.

The rest of the paper is organized as follows: The next section presents the methodology, Section 3 then sets out the computationally feasible estimator we use, and Section 4 conducts simulations comparing our methodology with the standard PEB approach in the literature. We then apply it in practice: Section 5 introduces the data, Section 6 describes the estimates of teacher VA using the two administrative data sets, and Section 7 presents the policy analysis. Section 8 concludes, considering the broader applicability of the approach.

2 Methodology

This section presents our methodology for estimating teacher value-added with reference to existing approaches in the literature.

2.1 Student Achievement and the Contribution of Teachers

We consider a standard model of student achievement in which education inputs (including the contribution of teachers) are additive in their effects. The achievement of a student i taught by teacher j in year t is written as:

$$\tilde{y}_{ijt} = X'_{ijt}\beta + \alpha_j + \epsilon_{ijt}, \quad i = 1, 2, \ldots, n_{jt}, \tag{2.1}$$

where $\tilde{y}_{ijt}$ is the student's observed test score (to be contrasted with y_{ijt} below, purged of covariates), and X_{ijt} are observed characteristics of the student (demographics, past academic performance, and family background) and the teacher (including her experience). Our parameter of interest, α_j, is the time-invariant teacher's contribution, or simply VA. We assume that teachers are each assigned to one class per year (with j's class size in year t being n_{jt}) and that conditional on X_{ijt}, the assignment is as good as random;[10] the error term ϵ_{ijt} is

[10]Rothstein (2017) and Chetty et al. (2017) discuss the validity of this assumption in the context of teacher value-added models. Our contribution to the value-added literature focuses on the empirical Bayes procedure,

assumed to be iid normal with variance σ_ϵ^2.[11]

The standard approach to estimating teacher VA starts from a regression that purges the effects of observed covariates from $\tilde{y}_{ijt}$. This leaves a noisy measure of the teacher's contribution, denoted

$$y_{ijt} = \alpha_j + \epsilon_{ijt}\,. \tag{2.2}$$

From here, several different estimators are available in order to estimate VA.

2.2 The Fixed Effect Estimator

Given (2.2), we can construct the maximum likelihood estimator (sometimes referred to as the fixed effect estimator) for the unobserved α_j. We will denote this by

$$y_j = \sum_t h_{jt} y_{jt} / \sum_t h_{jt} = \sum_t n_{jt} y_{jt} / \sum_t n_{jt}\ , \tag{2.3}$$

where $y_{jt} = \frac{1}{n_{jt}} \sum_{i=1}^{n_{jt}} y_{ijt}$ is the teacher-year specific sample average for teacher j in year t. Taking a weighted average of $\{y_{jt}\}$ across all the classes taught by teacher j over time, using weights $h_{jt} \equiv n_{jt}/\sigma_\epsilon^2$, gives the teacher fixed effect ('FE') for that teacher in (2.3). Together with the assumption that ϵ_{ijt} follows a normal distribution, model (2.2) then implies that the teacher FE has the following distribution:

$$y_j \sim \mathcal{N}(\alpha_j, \sigma_\epsilon^2 / \sum_t n_{jt}). \tag{2.4}$$

Given the expression for the variance, if the total sample $\sum_t n_{jt} \to \infty$ in the denominator, then fixed effect y_j converges to the true teacher VA, α_j, in probability, and so is a consistent estimator for the desired object. In practice, however, the VA literature does not use the fixed effect estimator, primarily because of finite sample considerations. These imply that

rather than gauging bias in value-added measures due to potential non-random assignment.

[11] This normality assumption is made to simplify the following discussion. It is not necessary. In Appendix A, we present a general framework demonstrating the nonparametric identification of the error distribution.

the fixed effect estimator is a noisy estimator, especially for teachers beginning their careers. Shrinkage estimators (considered next) account for this feature.[12]

2.3 The Parametric Empirical Bayes Estimator

The current state-of-art estimator for VA – the parametric empirical Bayes ('PEB') estimator introduced first by Kane and Staiger (2008) and further developed by Chetty et al. (2014a) – responds to the finite-sample considerations just referred to. It does so by leveraging the insight that if the teacher effect follows a normal distribution, then it is possible to modify poor-quality estimates for some teachers based on observations for other teachers.

The PEB estimator is the feasible version of the parametric Bayes ('PB') estimator, which is the minimizer of the Bayes risk, $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}(\delta_j - \alpha_j)^2]$, given (2.4) as well as the parametric assumption that the VA for all teachers is an independent and identically distributed draw from a normal distribution with mean zero and variance σ_α^2. The latter estimator takes the following form:

$$\delta_j^{PB} = y_j \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2 / \sum_t n_{jt}} \ . \tag{2.5}$$

Several remarks about the parametric Bayes estimator in (2.5) are due:

1. When $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$, the posterior distribution of α_j conditional on observing the teacher's performance y_j also follows a normal distribution, given by $\alpha_j | y_j \sim \mathcal{N}(y_j \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2/\sum_t n_{jt}}, \frac{\sigma_\alpha^2 \sigma_\epsilon^2/\sum_t n_{jt}}{\sigma_\alpha^2 + \sigma_\epsilon^2/\sum_t n_{jt}})$, where the posterior mean of α_j given the FE y_j is the best linear predictor of α_j.

2. The 'shrinkage' factor, $\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2/\sum_t n_{jt}}$ is always smaller than 1, which implies that the parametric Bayes estimator δ_j^{PB} shrinks the fixed effect estimator y_j towards zero. Given this, the direction of shrinkage only depends on the sign of y_j (not its magnitude).

[12] They can be assessed by their distance from the true VA quantity, α_j, using a common distance measure – the so-called $\mathcal{L}_2$ loss. This is denoted $L(\hat{\boldsymbol{\delta}}, \boldsymbol{\alpha}) \equiv \frac{1}{J}\sum_{j=1}^{J}(\hat{\delta}_j - \alpha_j)^2$, where $\hat{\delta}_j$ is some estimator of true VA, α_j.

3. The shrinkage factor is the same for all teachers with a given total sample size $n_j \equiv \sum_t n_{jt}$ (summing across all classrooms in all relevant time periods): the bigger the total sample size, the closer the shrinkage factor is to 1.

4. There is built-in symmetry in the Bayes estimator in the sense that, for individual teachers who have the same total sample size n_j, the amount of shrinkage imposed on y_j only depends on its absolute value. Thus teachers with both very large y_j (for example, teachers in the right tail, who have large positive fixed effects) and with small y_j (e.g., left-tail teachers) are shrunk towards zero by the same magnitude as long as their overall sample sizes and absolute magnitudes are the same.

5. The estimator δ_j^{PB} is infeasible since it involves unknown parameters $(\sigma_\alpha^2, \sigma_\epsilon^2)$. The empirical counterpart to δ_j^{PB}, the PEB estimator defined as $\delta_j^{PEB} = y_j \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2/n_j}$, replaces these unknown parameters with their consistent estimates, either through maximum likelihood or method of moments.

These observations serve to highlight two key features of the PEB estimator. First, by construction, the estimator δ_j^{PEB} is linear in y_j; hence it scales the fixed effect estimator in the same symmetric fashion towards zero for all teachers who have the same overall sample size n_j. Second, for teachers who have very large n_j relative to σ_ϵ^2, the PEB estimator is almost the same as the fixed effect y_j. This confirms the intuition from above that if n_j is large, the fixed effect will provide an accurate estimator for the true value-added, so the shrinkage estimator δ_j leaves it relatively unmodified. In contrast, teachers with a smaller total sample size (new teachers, for example) receive greater shrinkage towards zero.

The PEB estimator has a simple linear form and is easy to compute, which helps to account for its popularity in the literature. Crucially, however, it relies on the parametric assumption that true teacher VA, α_j, follows a normal distribution. If this assumption is misplaced, the quality of the shrinkage estimator may deteriorate,[13] perhaps significantly,

[13]See Bonhomme and Weidner (2019) for a discussion of the robustness properties of the EB estimator when the normality is locally misspecified.

and it raises the further possibility that one might be able to find an alternative estimator that has a smaller Bayes risk.

2.4 The Nonparametric Bayes Estimator

Next we show that, for model (2.2), the distribution of teacher VA is nonparametrically identified (Theorem 1 below). This result implies that the data contain enough information about the distribution of the true VA measure, and the normality assumption applied to unobserved teacher quality involves an over-parameterization. In the theorem, we continue to assume a normally distributed error, although doing so is not necessary (as we show in Appendix A, drawing on Kotlarski (1967)). Imposing normality will, however, be convenient for estimation purposes. We then introduce the nonparametric Bayes ('NPB') estimator for teacher VA (see Theorem 2). This estimator involves the unknown VA distribution, which we replace by an empirical counterpart (see Section 3), referring to that as the nonparametric empirical Bayes ('NPEB') estimator.

Theorem 1 *Consider the model $y_{ijt} = \alpha_j + \epsilon_{ijt}$, with $\epsilon_{ijt} \sim_{iid} \mathcal{N}(0, \sigma_\epsilon^2)$. If α_j is independent of ϵ_{ijt} for all i and t, and α_j follows some probability distribution F, then F is nonparametrically identified.*

Proof. Under the assumption that α_j and ϵ_{ijt} are independent random variables for all i and t, we have for any $s \in \mathbb{R}$,

$$\begin{aligned}
\phi_{y_{ijt}}(s) &= \int e^{\mathtt{i}sy} \int \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{(y-\alpha)^2}{2\sigma_\epsilon^2}} dF(\alpha) dy \\
&= \int e^{\mathtt{i}sz} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{z^2}{2\sigma_\epsilon^2}} dz \int e^{\mathtt{i}s\alpha} dF(\alpha) \\
&= e^{-\sigma_\epsilon^2 s^2/2} \phi_\alpha(s)\,,
\end{aligned}$$

where $\phi_X(\cdot)$

is the characteristic function of random variable X and $\mathtt{i}$ denotes the imaginary unit (as distinct from the student index i). Since we observe y_{ijt}, the characteristic function

$\phi_\alpha(t)$ is identified from the data for all $s \in \mathbb{R}$. Given the one-to-one mapping from the characteristic function to the distribution function of a random variable, the distribution F is nonparametrically identified. ■

Next, we present the nonparametric Bayes estimator, using teacher fixed effects (y_j) and the model (2.4) as inputs:

Theorem 2 *Given the model $y_j = \alpha_j + \nu_j$, with α_j being independent from ν_j, $\alpha_j \sim F$ and $\nu_j \sim \mathcal{N}(0, \sigma_\epsilon^2/n_j)$, then the estimator of α_j that minimizes the Bayes risk under $\mathcal{L}_2$ loss, $\tilde{\alpha}_j$, takes the form*

$$\delta_j^{NPB} = \frac{\int \alpha \varphi_j(y_j - \alpha) dF(\alpha)}{\int \varphi_j(y_j - \alpha) dF(\alpha)}, \tag{2.6}$$

with $\varphi_j(\cdot)$ being the density function of the normal distribution with mean zero and variance σ_ϵ^2/n_j. The estimator can be further simplified to

$$\delta_j^{NPB} = y_j + \frac{\sigma_\epsilon^2}{n_j} \frac{\partial}{\partial y} \log g_j(y)|_{y=y_j}\,, \tag{2.7}$$

with $g_j(\cdot)$ being the marginal density of y_j.

Proof. The first part of proof follows from the fact that the minimizer of the Bayes risk under $\mathcal{L}_2$ loss is the posterior mean of α_j conditional on y_j. For the second part, since

$$g_j(y) = \int \varphi_j(y - \alpha) dF(\alpha),$$

then straightforward calculations show that

$$\frac{\sigma_\epsilon^2}{n_j} \frac{\partial}{\partial y} \log g_j(y)|_{y=y_j} = \frac{\int (\alpha - y_j) \varphi_j(y_j - \alpha) dF(\alpha)}{\int \varphi_j(y_j - \alpha) dF(\alpha)} = \mathbb{E}[\alpha | y_j] - y_j\ .$$

Therefore,

$$\delta_j^{NPB} = \mathbb{E}[\alpha | y_j] = y_j + \frac{\sigma_\epsilon^2}{n_j} \frac{\partial}{\partial y} \log g_j(y)|_{y=y_j}\ .$$

■

Theorem 2 presents two expressions for the nonparametric Bayes estimator. The expression in (2.7) is known in the literature as "Tweedie's formula" (see Efron (2011)). Several remarks about it are due, compared with the parametric Bayes estimator in (2.5):

1. The parametric Bayes estimator is a special case of the nonparametric Bayes estimator.[14]

2. Tweedie's formula retains the feature that if σ_ϵ^2/n_j is very small, then the nonparametric estimator δ_j^{NPB} will not deviate much from the fixed effect estimator y_j.

3. In general, for any distribution F other than the normal distribution, the quantity $\frac{\partial}{\partial y}\log g_j(y)|_{y=y_j}$ in Theorem 2 introduces a non-linearity into the shrinkage rule with respect to y_j.

4. Unlike the parametric Bayes formula, Tweedie's formula does not automatically 'shrink' the fixed effect estimator towards zero. Inspection of the formula shows that the fixed effect and the adjustment factor are additively separable, and the adjustment factor can be positive or negative. The direction and magnitude of the shrinkage depend on the distribution of latent teacher quality and also the magnitude of the fixed effect estimate y_j, and hence is likely to be context-specific. (An illustrative example is provided below to give intuition.)

Given these observations, it is worth highlighting an important contrast: unlike the PEB estimator, individuals with higher variances will not necessarily shrink the most towards zero under Tweedie's formula. This feature turns out to be empirically relevant. Suppose a new teacher j who has a small total sample size n_j relative to σ_ϵ^2 happens to have a high y_j – for example, a teacher who performs very promisingly in her early years in the school system. Under the parametric shrinkage rule, this teacher will be heavily discounted in that her VA measure will shrink significantly towards zero. This is simply because the normal

[14]Specifically, when $F = \mathcal{N}(0, \sigma_\alpha^2)$, we have $g_j(\cdot)$ becoming the density function of a normal distribution with mean zero and variance $\sigma_\alpha^2 + \sigma_\epsilon^2/n_j$ and thus $\frac{\partial}{\partial y}\log g_j(y)|_{y=y_j} = -\frac{y_j}{\sigma_\alpha^2+\sigma_\epsilon^2/n_j}$ and $\delta_j^{NPB} = \delta_j^{PB} = y_j\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2/n_j}$.

distribution, which dictates that there must be a thin tail, deems it unlikely for this teacher to have a high VA α_j; instead, the large y_j arises purely by chance due to the associated high variance. In contrast, under the nonparametric shrinkage rule, depending on the features of the distribution of true value-added, her VA estimate may remain very close to her estimated fixed effect, y_j, and conceivably be even higher.[15]

Example: We now provide an example to illustrate the ways in which shrinkage may operate under the nonparametric Bayes estimator as compared with the PEB estimator, depending on the underlying distribution of teacher VA. Specifically, we consider the Bayes estimator under two different assumptions about the true teacher VA distribution. In the first case, assume $\alpha_j \sim \mathcal{N}(0, 0.05)$.[16] Here, we draw on the fact that the nonparametric Bayes and parametric Bayes estimators will coincide. In the second case, suppose the true teacher VA has distribution F, which takes the following form:

$$F = 0.98\mathcal{N}(0, \theta_1) + 0.01\mathcal{N}(-1, \theta_2) + 0.01\mathcal{N}(1, \theta_3)\,. \tag{2.8}$$

This mixed normal distribution has the built-in feature that at both tails, there is a small probability mass concentrated around the values -1 and 1, while the majority of the probability mass follows a normal distribution centered at zero; as such, this will help to highlight the operation of nonlinear shrinkage. Here, we calibrate the above parameters $(\theta_1, \theta_2, \theta_3)$ such that F has the same mean and variance as in the first case. In both, we set $\sigma_\epsilon^2 = 0.25$, which is roughly the same as in the North Carolina mathematics score data.

Figure 1 compares the amount of shrinkage for the parametric and nonparametric Bayes estimators, respectively. It does so for a hypothetical teacher j with total sample size $n_j = 20$ with a fixed effect estimate, y_j, in the range of $[-2, 2]$. As noted, the parametric Bayes

[15] A similar thought experiment can be conducted for a teacher who shows very poor observed performance (that is, who has a large negative y_j) and has a relatively small total sample size. Under the parametric shrinkage rule, this teacher would be regarded as more similar to the mean quality teacher, while the nonparametric rule does not adjust in a mechanical way, but allows for a range of possibilities that are informed by the data.

[16] Looking ahead, this will roughly coincide with the parametric empirical Bayes estimates obtained using the North Carolina data on mathematics scores.

estimator shrinks estimates linearly and symmetrically toward zero, with the direction of the shrinkage depending only on the sign of y and not its magnitude. In contrast, the nonparametric Bayes estimator (in equation (2.7)) is a nonlinear function of y and the direction of shrinkage depends not only on the sign of y but also its magnitude.

The figure makes clear that differences in the amount of shrinkage between the two estimators are especially pronounced at the mass points of 1 and -1. On one hand, the parametric Bayes estimator does not account for the presence of these mass points (as they are assumed away by normality) and so shrinks them towards zero. The NPB, on the other hand, adapts to the true underlying distribution – we discuss why below – and pulls fixed effects nearby to these two mass points, accounting for the non-negligible probability that the true quality of some teachers may take values of these mass points.

The nonparametric Bayes estimator δ_j^{NPB}, specified in equation (2.6), is infeasible in practice since it involves the unknown quantity σ_ϵ^2 and the distribution F. We develop a feasible version of the nonparametric estimator next.

3 A Feasible Nonparametric Bayes Estimator

This section sets out the feasible version of the nonparametric Bayes estimator, which we refer to as the nonparametric empirical Bayes (or NPEB) estimator, the primary focus for the remainder of the analysis.

By way of overview, we take an approach that shares the same spirit as the parametric empirical Bayes method. For the unknown parameter σ_ϵ^2, we use its maximum likelihood estimator (see Appendix D), while we estimate the distribution F directly from the data (instead of assuming it belongs to a parametric distribution family). This then yields the NPEB estimator (see equation (3.2) below).

3.1 Nonparametric Maximum Likelihood Estimation of the Distribution F

Our approach to recovering the teacher VA distribution F is based on methods proposed in Jiang and Zhang (2009), Koenker and Mizera (2014) and Gu and Koenker (2017b). Those papers set out a general framework for estimating unobserved heterogeneity in cross-sectional and longitudinal data settings without imposing any parametric assumptions on the unobserved heterogeneity, drawing on the seminal contribution by Robbins (1956), referenced above. As the methodology fits many contemporary 'Big Data' applications, the revival in the use of nonparametric empirical Bayes methods for large-scale inference is natural,[17] reflected in several recent applications in statistics and economics.[18] The teacher VA application fits into this general framework well, as teacher quality can be thought of as unobserved heterogeneity in a model of test scores that accounts for test score variation unexplained after controlling for all observed heterogeneity through covariates X_{ijt}.

We denote the distribution of α_j in a general way as F, rather than assuming $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$. The distribution F is not observed by the researcher, but can be estimated nonparametrically from the data via the following optimization:

$$\hat{F} \equiv \underset{F \in \mathcal{F}}{\operatorname{argmax}} \left\{ \sum_{j=1}^{J} \log \int \varphi_j(y_j - \alpha) dF(\alpha) \right\} \tag{3.1}$$

where φ_j is a normal density with mean zero and variance σ_ϵ^2/n_j, as in (2.4), and the space $\mathcal{F}$ is the set of all probability distributions on $\mathbb{R}$. The resulting $\hat{F}$ is the nonparametric maximum likelihood estimator (hereafter 'NPMLE') for F.

Kiefer and Wolfowitz (1956) established consistency of the NPMLE for the mixing distribution F. A viable computational strategy for the estimator was not available until the

[17] See Efron (2010) for a survey. A recent simulation comparison across different machine learning methods is provided in Abadie and Kasy (2019), highlighting the advantages of the nonparametric empirical Bayes method for high-dimensional settings.

[18] These include predicting baseball batting averages (Gu and Koenker, 2017a), studying earning dynamics (Gu and Koenker, 2017b), analyzing treatment effect heterogeneity (Gu and Shen, 2017), and estimating the impact of neighborhood effects on intergenerational mobility (Abadie and Kasy, 2019).

appearance of the EM algorithm by Laird (1978), and the EM algorithm has remained the standard approach for its computation ever since.[19] However, EM has notoriously slow convergence in nonparametric EB applications, especially with large data sets, and this fact has inhibited the widespread implementation of the NPMLE.

Koenker and Mizera (2014) recently proposed an alternative computational method for the NPMLE that circumvents these issues. They show that for a broad class of mixture problems, the Kiefer-Wolfowitz estimator can be formulated as a convex optimization problem and solved efficiently by modern interior point methods. Quicker, more accurate computation of the NPMLE in turn opens up a much wider range of applications of the method for models with heterogeneity.[20] Here, the large-scale data involved in teacher VA estimation make it likely to benefit from the scalability of the new computational method.

The difficulty in estimating equation (3.1), as pointed out in Koenker and Mizera (2014), is that F is an infinite dimensional object and so there is an infinite number of constraints. To make computation feasible and maintain the convexity of the problem, they propose a finite-dimensional convex approximation. Formally, let M be a positive integer and let $\mathcal{F}_M$ be the class of probability distribution functions supported on M grid points given by $\min_j\{y_j\} < \alpha_1 < \alpha_2 < \cdots < \alpha_M < \max_j\{y_j\}$. The NPMLE is then defined to be the maximizer of equation (3.1), replacing $\mathcal{F}$ by $\mathcal{F}_M$. The resulting NPMLE $\hat{F}$ thus takes the form of a discrete distribution.[21] Note that the size of the grid M is not a tuning parameter. When M is reasonably large, increasing it further does not improve the likelihood; Dicker and Zhao (2016) show that taking M to be roughly the square-root of the sample size renders a good approximation.

[19] Heckman and Singer (1984) constitutes an influential econometric application.

[20] Models with unobserved heterogeneity beyond the normal mixture are discussed in Koenker and Gu (2017).

[21] For the data sets we apply this method to, the sample size is around 35,000 for the North Carolina data and 11,000 for the LAUSD data, and in both cases, we take $M = 5000$.

3.2 The Plug-in Nonparametric Empirical Bayes Estimator for VA

With the NPMLE $\hat{F}$, we can construct the NPEB VA estimator as:

$$\delta_j^{NPEB} = \frac{\int \alpha \varphi_j(y_j - \alpha) d\hat{F}(\alpha)}{\int \varphi_j(y_j - \alpha) d\hat{F}(\alpha)}. \tag{3.2}$$

The estimator defined in equation (3.2) is the feasible version of the posterior mean defined in equation (2.6) in Theorem 2. Once we obtain the NPMLE for F based on (3.1), evaluating (3.2) only involves matrix operations, since $\hat{F}$ already takes a discrete form (as noted).

We focus on constructing a feasible version of the NPB estimator based on (2.6) directly, rather than its equivalent reformulation (2.7) in Theorem 2, for two reasons. First, equation (2.7) suggests that the nonparametric Bayes estimator δ_j^{NPB} does not depend on the teacher quality distribution F directly, but rather on the marginal density of the fixed effect estimator. Therefore, in principle we could focus on constructing a feasible estimator for the marginal density. In practice, this becomes challenging when individual teachers have heterogeneous variances.[22]

Second, kernel-based estimators for the marginal density do not incorporate the model information that the fixed effect estimator is induced by an underlying normal mixture model; hence the resulting shrinkage estimator may lose some important properties of the NPB estimator, such as monotonicity with respect to y_j for fixed variances.[23] In contrast, both the construction of the NPMLE of F and the NPEB in (3.2) make use of the mixture model structure, and so automatically satisfy the monotonicity property.[24]

[22] Brown and Greenshtein (2009) proposes a kernel method to estimate the marginal density of y_j directly when variances of y_j are all the same – when all teachers have the same associated sample sizes, for example. When variances are homogeneous, the kernel estimator for the marginal density is easy to construct since we have J independent and identically distributed observations $(y_1, \ldots, y_J)$ from this marginal density. Yet when individual teachers have heterogeneous variances, it is difficult to apply these methods to construct (2.7), given the observations $(y_1, \ldots, y_J)$ are no longer identically distributed.

[23] See Koenker and Mizera (2014) for a monotonized version of the Brown and Greenshtein (2009) estimator.

[24] We can motivate the proposed estimator (3.2) on the grounds that it makes use of both the information from the data (to learn about F) and from the model (using the normal mixture structure). Saha and Guntuboyina (forthcoming) recently showed that the NPEB estimator δ_j^{NPEB} constructed via the NPMLE

The newly proposed estimator δ_j^{NPEB} has the potential to improve on the linear PEB estimator (in the sense of having a smaller average squared error) when the underlying distribution F cannot be approximated well by the normal distribution. The magnitude of this change can be evaluated through simulations, where the distribution F used to generate the data is known. We conduct such illustrative simulations next.

4 Simulations

In this section, we use simulations to compare the performance of three estimators: the fixed effects estimator, the PEB estimator, and our NPEB estimator. We do so relative to a benchmark – infeasible in practice – in which the researcher knows the true underlying teacher quality distribution, and can therefore use the optimal Bayes rule.

We consider the performance of these candidate estimators on the basis of their mean squared error under three candidate distributions – normal, mixed normal, and chi-squared. Simulated data are generated using equation (2.2) for 10,000 individual teachers with $\epsilon_{ijt} \sim \mathcal{N}(0, 0.25)$, set to mimic what we estimate using data later. For comparison, we consider both a homogeneous class size case where every teacher has a class size of 20, and a heterogeneous class size case where class size is drawn randomly from the set $\{20, 40\}$ with equal probability.

The Teacher Quality Distribution is Normal: Table 1(a) displays the simulation results when teacher quality is normally distributed according to $F \sim \mathcal{N}(0, 0.05)$. Here, the normality assumption already built into the EB estimator is correct, and so it performs identically to the infeasible estimator where the distribution is known. We see that the PEB estimator improves on the mean squared error of the fixed effects estimator substantially. At the same time, it only outperforms our NPEB estimator by a very small margin: mean squared error is less than one percent higher under NPEB – striking given the NPEB does

of F performs similarly to the infeasible NPB estimator δ_j^{PEB} defined in (2.6). The proposed estimator achieves this satisfactory approximation to the infeasible estimator, despite of the well-known fact that the NPMLE of F has a slow convergence rate (Fan, 1991), because the nonparametric Bayes rule is a *smooth* functional of F, which can be estimated at a much better rate than the distribution F itself.

not make *any* parametric assumption about the distribution F.

The Teacher Quality Distribution is Non-Normal: Tables 1(b) and 1(c) show simulation results when teacher quality is not normally distributed. Specifically, in Table 1(b), true teacher quality follows the mixed normal distribution $F \sim 0.98\mathcal{N}(0, 0.03)+0.01\mathcal{N}(-1, 0.03)+0.01\mathcal{N}(1, 0.03)$, as in equation (2.8), while in Table 1(c), true teacher quality follows $F \sim \chi^2_1$.[25]

Here, it is clear that our NPEB estimator outperforms *both* the PEB estimator and the fixed effects estimator substantially in terms of mean squared error. Even more noteworthy is the fact that our NPEB estimator achieves a mean squared error very close to that of the infeasible estimator (which assumes the true distribution is known): our NPEB estimator has mean squared error less than one percent higher than the infeasible estimator for both distributions considered. In contrast, the mean squared errors of the PEB and fixed effect estimator are 15-60 percent higher. Of note, the PEB estimator performs worse as the underlying distribution deviates further from normality; thus its efficacy is particularly poor for the chi-squared distribution, as shown in Table 1(c).

In sum, as these simulations indicate, our proposed approach has an appealing versatility in that it can adapt to any distribution. Despite being a nonparametric method, it performs similarly to the PEB estimator when the true distribution is normal. When the true distribution is not normal, our approach is able to outperform the PEB estimator by a substantial margin, performing almost as if the true distribution were known. Given there is no *a priori* reason to believe that teacher quality follows some specific distribution, the simulation evidence we have presented buttresses the view that our method can adapt to, and help recover, any underlying distribution of teacher quality. Next, we take the method to observational data.

[25]For comparability, the mixed normal distribution has the same mean and variance as the normal distribution in Table 1(a).

5 Data

Our data are drawn from two administrative data sets, each providing detailed information about students and teachers, including enrollment history, test scores and teacher assignments. The data from both sources cover similar time periods, grades, and demographic information, although there are important differences. For that reason, we discuss each data source separately; a more detailed description is provided in Appendix B.

North Carolina: Our first administrative data set covers all public school students in North Carolina for third through fifth grade – specifically, third grade from 1996-97 to 2008-09 and for fourth and fifth grades from 1996-97 to 2010-11.[26] These data cover around 1.85 million students with 4.5 million student-year observations. We also have detailed demographic information including parental education (1996-97 through 2005-06), economically disadvantaged status (1998-99 through 2010-11), ethnicity, gender, limited English status, disability status, academically gifted status, and an indicator for grade repetition.

We make several restrictions to construct the sample used for estimating teacher VA, following Clotfelter et al. (2006) and subsequent research using North Carolina data. Specifically, we require that all students are matched to a teacher, and that students have a valid lagged test score in the relevant subject. After the sample restrictions, our final sample consists of approximately 2.7 million student-year observations, covering 1.4 million students and 35,000 teachers.

Table 2 provides summary statistics for the main variables used in calculating VA. Column (1) reports these for the entire North Carolina sample, and column (2) for the VA analysis data set. While the sample restrictions eliminate approximately forty percent of the observations, we see only minor differences between the two samples, with the VA sample showing slightly higher performance levels and being drawn from moderately higher socioeconomic backgrounds on average.

[26] Our analysis is restricted to students in those grades since our data set records the test proctor, and the test proctor is typically the teacher who taught the students throughout the year in these grades. Data for third grade ceases after 2008-09 because the third grade pretest was discontinued after that year.

Los Angeles Unified School District: Our second data source comes from the Los Angeles Unified School District (LAUSD). The data set spans third grade from 2003-04 to 2012-13 and fourth and fifth grade from 2003-04 to 2012-13 and 2015-16 to 2016-17.[27] It covers roughly 800,000 students with 1.7 million student-year observations. Detailed demographic data include economically disadvantaged status, ethnicity, gender, age, limited English status, and a grade repetition indicator, and parental education (missing for thirty percent of sample).

Similar to the North Carolina data set, we make several sample restrictions, dropping students who cannot be matched to a classroom teacher, and students who do not have a valid lagged test score in the relevant subject. Our VA sample for LA consists of 1.3 million student-year observations, covering roughly 660,000 students and 11,000 teachers.

Columns (3) and (4) of Table 2 provides summary statistics for the LAUSD data. Column (3) reports these for the entire sample, while column (4) does so for the VA analysis data set. Similar to the North Carolina case, we find that our VA sample is moderately positively selected, with student test scores being about 0.06 standard deviations higher than the full sample.

Comparing the two administrative data sets, clear differences in samples become apparent. While North Carolina has a majority-white student body with a large black minority, the LAUSD is majority-Hispanic. The LAUSD sample is also drawn from students with significantly more disadvantaged backgrounds, with students being almost twice as likely to be free or reduced price lunch-eligible and nearly three times as likely to come from a household where parents are high school dropouts. These differences may attract particular teachers, potentially giving rise to a different underlying distribution of teachers in these two settings.

[27] Third grade data are missing for 2013-14 and 2014-15 due to a change in the statewide testing regime that occurred in 2013-14, which resulted in no test score data that year and also eliminated the second grade test thereafter. As lagged test scores are required when computing value-added, we drop academic years 2013-14 and 2014-15 from the dataset as well as third grade after 2012-13.

6 Results

This section reports estimates of teacher VA using our proposed NPEB methodology, alongside those using the PEB approach. We describe results for North Carolina and the LAUSD separately, highlighting differences in the estimated teacher quality distributions. Then we discuss how these differences will be relevant for the policy analysis in Section 7, and also compare the out-of-sample performance of our NPEB estimator relative to the PEB estimator.

6.1 VA Estimates

North Carolina: Figure 2(a) displays a boxplot of teacher fixed effects for teachers who appear once, twice, three times or more than three times, respectively, in our North Carolina data set. This shows that teachers appearing for more periods – typically, more experienced teachers – exhibit less dispersion as the fixed effect for these teachers is estimated with a larger effective sample size than teachers appearing less frequently. At the same time, the average fixed effect is similar, regardless of how often teachers appear in the data (conditional on teacher experience). Bayesian shrinkage is then applied to these fixed effects, serving to shrink fixed effect estimates for teachers with small sample sizes back towards the mean. The boxplot in Figure 2(c) shows the magnitude of shrinkage applied by our NPEB estimator. As expected, only small amounts of shrinkage arise for teachers with more than three years of data, while teachers who appear less frequently and thus have smaller effective sample sizes are shrunk towards zero in a more pronounced way.

The estimated teacher VA distribution used for our NPEB (estimated using equation (3.1)) is shown in Figure 3(a). At first glance, it appears to be approximately normal, although policymakers are particularly interested in the tails of the distribution. Given that focus, Figure 3(b) takes the cube root of the distribution in order to enhance the tails. Having done so, we can see that the teacher quality distribution exhibits a fat right tail relative to the Gaussian distribution. In light of this, one might anticipate that applying our

NPEB methodology in North Carolina would mostly agree with the PEB method (which assumes normality), except for the right tail.

Next, we compare our estimator with the PEB estimator. To do so, we start by estimating the distribution of teacher VA under the normality assumption using maximum likelihood.[28] Based on mathematics scores from all North Carolina school districts, we find that teacher VA is distributed $\mathcal{N}(0, 0.047)$, implying a standard deviation of 0.217, and that $\hat{\sigma}_\epsilon^2 = 0.248$. Figure 4(a) then compares the VA estimates obtained from the PEB and NPEB estimators for representative teachers with a class size of twenty students but who may have different fixed effect estimates (shown on the horizontal axis). In accord with expectations, teacher VA is nearly identical across the two methods for teachers with fixed effect estimates below 95^{th} percentile of its distribution (indicated by the vertical dashed line to the right) of the figure. There is some disagreement between the two methodologies in the right tail, with the NPEB estimator not shrinking teachers who have a fixed effect in the far right tail as markedly as the PEB estimator. This can be seen with the NPEB estimates (shown by the curved line) being substantially above those estimated using the PEB methodology (the dashed upward-sloping straight line) and closer to the 45 degree line. Intuitively, this misalignment occurs because our nonparametric methodology finds a distribution with a fatter right tail than the normal distribution assumed under the parametric method, and so the NPEB estimator shrinks teachers less aggressively if they have noticeably large fixed effect estimates.

LAUSD: Figures 2(b) and 2(d) present information for LAUSD – boxplots of teacher fixed effects and the magnitude of shrinkage that our NPEB estimator applies according to how often teachers appear in the data. These figures are similar to those for North Carolina, although the LAUSD exhibits a higher variance in teacher fixed effects, which leads to higher dispersion in VA in the district.

Our estimated teacher VA distribution using NPEB is shown in Figure 3(c). It is more dispersed and skewed than the corresponding distribution for North Carolina. Enhancing

[28] This is the Chetty et al. (2014a) no-drift estimator. Maximum likelihood is used as it is the most efficient estimator: results are similar if we use method of moments instead.

the tails in Figure 3(d) draws attention to other compelling differences: the VA distribution has a much thinner left than right tail. Our NPEB methodology should therefore mostly agree with the PEB method, except in the two tails.

Using parametric EB, we find teacher VA is distributed as $\mathcal{N}(0, 0.0977)$, implying a standard deviation of 0.3126, and that $\hat{\sigma}_\epsilon^2 = 0.2596$, which is considerably higher than the variance we found in North Carolina. Comparing the VA estimates obtained from the two empirical Bayes estimators, again for representative teachers with a fixed class size of twenty students, Figure 4(b) indicates that teacher VA is very similar for teachers with their fixed effect estimates in the 5-95th percentile of its distribution, but deviates in both tails, and especially in the left tail. Our nonparametric method shrinks them far more strongly back toward the mean when the fixed effect estimates take a value below the 5th quantile while shrinking them less aggressively when the fixed effect estimates take a value above the 95th quantile, in order to adapt to the skewness of the distribution.

Implications for Policy: These differences in the tails of the VA distribution relative to a normal distribution can drive large differences in policy calculations between the two empirical Bayes methodologies. On the one hand, the normality assumption is particularly misspecified in the right tail for North Carolina. Given the set of teachers affected by interventions targeting high-VA teachers, we expect that the gains from such policies – e.g., retention bonuses – will be *understated* in the PEB methodology since it shrinks high-VA teachers back towards the mean too strongly.

On the other hand, the normality assumption is misspecified in the left tail for the LAUSD, which is important for policies targeting low-VA teachers, such as teacher release policies. Here, the PEB methodology is likely to *overstate* the benefit of such policies since it does not shrink these low-VA teachers sufficiently back toward the mean.

Teacher Rankings: A natural question arises whether the choice of estimator influences the *ranking* of teachers. Consider misclassification rates of teachers in the bottom five percent

in terms of both Type I and Type II errors.[29] With homogeneous class sizes, the PEB and NPEB estimators are both monotone functions of teacher fixed effects (Guarino et al., 2015; Bitler et al., 2019), guaranteeing that teacher rankings (and thus misclassification rates) are identical across the fixed effect estimator and PEB and NPEB estimators. Once we allow class size to be different, however, the EB and NPEB estimators are no longer order-preserving with respect to teacher fixed effects, raising the possibility that misclassification rates may differ depending on the estimator used, in turn making it unclear which estimator should (in principle) be used to determine *whom* to replace.[30]

In practice, we find that the choice of empirical Bayes method has little appreciable impact on teacher rankings. In both data sets, a very small fraction of teachers ranked in the bottom five percent of the teacher quality distribution under PEB is *not* ranked in the bottom five percent according to NPEB. (The exact numbers of teachers ranked in the bottom five percent under PEB but not under NPEB for our two datasets are: 34 out of 1753 for North Carolina and 33 out of 554 for the LAUSD, respectively.) This is in line with our simulation results, which reveal nearly identical misclassification rates across the three estimators – see Tables F.1(a)-F.1(c).

6.2 Out-of-Sample Predictions

Given our estimates of VA described above, we now evaluate the performance of our NPEB estimator relative to the fixed effect and PEB estimators. A natural way of evaluating the estimators is on the basis of their ability to predict future outcomes. For instance, suppose that school boards or policymakers observe a teacher's past performance and wish to predict future outcomes for that teacher. We can measure the performance of this prediction via the squared error distance, $(y_{j,t+1} - \hat{y}_{j,t+1})^2$, where $y_{j,t+1}$ is the true outcome of teacher j in period $t+1$ and $\hat{y}_{j,t+1}$ is its predictor, utilizing all past information relating to her teaching

[29] In this context, a Type I error occurs when a teacher is ranked below 5% while her true quality ranking is above 5%; conversely, a Type II error occurs when a teacher is ranked above 5% when her true quality is below 5%.

[30] For instance, Mehta (2019) finds that policymakers should not use the empirical Bayes correction under certain systematic relationships between teacher quality and class size.

performance, starting when the teacher first appeared in the sample up until the t-th period.

This prediction exercise faces one inherent difficulty in that the class sizes of teachers in period $t+1$ differ. Thus, even if we had a perfect estimator for a teacher's quality α_j, since $y_{i,t+1} \sim \mathcal{N}(\alpha_j, \sigma_\epsilon^2/n_{j,t+1})$, the larger the class size, the less variability there would be in outcomes the following year, given by $y_{j,t+1}$, making teacher quality easier to predict for the corresponding teacher. To account for this, we use the following two measures of prediction accuracy proposed in Brown (2008): normalized mean squared error (NMSE) and total mean squared error (TMSE). NMSE is given by:

$$NMSE = \frac{1}{K} \sum_{j \in I} \left(n_{j,t+1} (y_{j,t+1} - \hat{y}_{j,t+1})^2 \right), \tag{6.1}$$

where I is the set of teachers whose performance is being predicted and K is the size of that set. The NMSE is the usual sum of squared errors with an adjustment term $n_{j,t+1}$ (where the adjustment term serves to scale back the contribution of teachers with large classes sizes, who have higher precision by construction, by the size of their class, $n_{j,t+1}$). Alternatively, TMSE is given by:

$$TMSE = \frac{1}{K} \sum_{j \in I} \left((y_{j,t+1} - \hat{y}_{j,t+1})^2 - \frac{\sigma_\epsilon^2}{n_{j,t+1}} \right). \tag{6.2}$$

Without the adjustment term $\frac{\sigma_\epsilon^2}{n_{j,t+1}}$, the quantity is just the usual sum of squared errors; the adjustment term is introduced to account for the effect of different class sizes on the variance.

Tables 3(a) and 3(b) report NMSE and TMSE using the North Carolina and LAUSD data for three different estimators: NPEB, parametric EB, and fixed effects. Each row in the table represents the number of prior years of teacher j's performance used to make the prediction.[31] The empirical Bayes methods (reported in the first two columns) substantially outperform the fixed effects method when only a few years of prior data are used, the extra

[31] To predict the performance of teacher j using t years of data, we restrict the sample to include teachers who appear $t+1$ times.

gain being substantial when information about a teacher is scarce. As more and more years of data become available, the gain from using empirical Bayes diminishes in comparison with the fixed effect estimator.

Comparing the two empirical Bayes estimators, the NPEB outperforms the PEB estimator under both prediction accuracy measures, except when only using one prior year of information with the LAUSD data set. When 2-5 years of data are used, the prediction performance of the NPEB estimator surpasses that of the PEB by the greatest margin. With many years of data, the NPEB continues to outperform the parametric EB, although both methods begin approaching the performance of the fixed effect estimator, given teacher-specific sample sizes have become large enough such that estimates are no longer shrunken materially. Of note, teacher tenure decisions are often made in practice during the time window when our nonparametric methodology outperforms that of the PEB by the greatest margin (namely using 2-5 years of data per teacher).[32]

7 Policy Analysis

This section performs policy calculations for policies that target the bottom and top of the teacher quality distributions, respectively. We pay particular attention to differences in policy calculations found using our nonparametric method relative to those using the PEB methodology. (Standard errors for our policy calculations are bootstrapped – see Appendix C for details.)

7.1 Policy Experiment I: Lay-off Policies

One policy recommendation that has gained considerable traction focuses on replacing poor-quality teachers with mean-quality teachers. The specific proposal made by Hanushek (2009, 2011) and further explored by Chetty et al. (2014b) involves releasing teachers in the

[32] Among the five most populous states, for instance, Texas, Pennsylvania and Florida award teacher tenure after three years, while California and New York award tenure after two and four years, respectively.

bottom 5% of the estimated VA distribution, replacing them with those who are of average quality. Here, we calculate the policy gains from (more generally) replacing the bottom q% of teachers, although we pay particular attention to the 'bottom 5%' cutoff, given its prominence both in the literature and policy debate.

Under parametric EB, the policy gain of replacing the bottom q% of teachers is calculated based on the assumption that teacher VA is normally distributed. Since teachers are assumed to be drawn from $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, replacement teachers who are at the mean of the distribution are those with $\alpha = 0$. In addition, since a one-unit increase in α (meaning VA) leads to a one SD test score gain, the marginal gain in test scores from such a policy, denoted as $MR(q)$, is:

$$MR(q) \equiv -\mathbb{E}[\alpha \mid \alpha < \Phi^{-1}(q)]\,, \tag{7.1}$$

where Φ is the cdf of $\mathcal{N}(0, \sigma_\alpha^2)$ and q is the cutoff percentage. Further, under the assumption that $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, it is straightforward to show that

$$MR(q) = -\frac{\int_{-\infty}^{\Phi^{-1}(q)} \alpha\varphi(\alpha)d\alpha}{\int_{-\infty}^{\Phi^{-1}(q)} \varphi(\alpha)d\alpha}\,, \tag{7.2}$$

with $\varphi(\alpha)$ being the pdf of $\mathcal{N}(0, \sigma_\alpha^2)$. Similarly, the total gain in test scores for a policy that deselects the bottom q% of teachers, denoted $TR(q)$, is:

$$TR(q) \equiv \mathbb{E}[\alpha \mathbb{1}\{\alpha \geq \Phi^{-1}(q)\}]\,, \tag{7.3}$$

where $\mathbb{1}\{\alpha \geq \Phi^{-1}(q)\}$ takes the value one if the teacher's quality is greater than the cutoff value $\Phi^{-1}(q)$ and zero otherwise. Again, under the assumption that $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, this can be written:

$$TR(q) = \int_{\Phi^{-1}(q)}^{+\infty} \alpha\varphi(\alpha)d\alpha\,. \tag{7.4}$$

The normality assumption embedded in the PEB estimator is likely misspecified, however, given the distributions we estimated nonparametrically in Section 6. Based on our estimated distributions, replacement teachers remain mean zero (i.e., $\alpha=0$) since VA is always centered. The marginal and total test score gains from a policy laying off the bottom q percent of teachers with our estimated distribution are instead given by:

$$MR(q) = -\frac{\int_{-\infty}^{F^{-1}(q)} \alpha f(\alpha) d\alpha}{\int_{-\infty}^{F^{-1}(q)} f(\alpha) d\alpha}\,, \tag{7.5}$$

$$TR(q) = \int_{F^{-1}(q)}^{+\infty} \alpha f(\alpha) d\alpha\,, \tag{7.6}$$

where $F(\alpha)$ and $f(\alpha)$ are the true cumulative probability distribution and density functions of unobserved teacher VA, respectively.

The distribution F is identified from the data (appealing to Theorem 1), so we can evaluate the policy under two distributions of VA: (i) the normal distribution with a standard deviation estimated from the data, and (ii) the distribution F estimated using the same data set. To begin, we assume that the policymaker can observe true teacher VA, an assumption we relax below.

Figure 5 compares the policy gains under the PEB methodology (where it is assumed that $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$) with those found using our NPEB method (where $\alpha \sim F$). For North Carolina, the PEB methodology overestimates policy gains, although not substantially. The picture is very different for LAUSD, however, with the PEB methodology overestimating the policy gains considerably (see the bottom two panels).

Table 4(a) quantifies these differences by reporting test score gains under the PEB methodology and our NPEB methodology respectively for a policy that releases the bottom q% of teachers for both North Carolina and the LAUSD. The bolded row indicates our benchmark policy that releases bottom-ventile (i.e., bottom-5 percent) VA teachers. Columns (1) and (2) indicate that the PEB method overstates the policy gains by about five percent relative to our NPEB method in North Carolina and so the normality assumption

appears reasonably well-founded in that context for policies targeting the left tail. In contrast, the PEB method overstates the policy gains by *twenty-four* percent in the LAUSD. While the differences in test score gains comparing the PEB and NPEB methods are highly statistically significant in both cases, the normality assumption is clearly far more misplaced in the LAUSD data, highlighting the benefit of using the flexible method we propose. The method adapts to the underlying distribution, which is unknown *a priori*.

Accounting for the Fact that VA is Estimated: The above policy analysis is based on the presumption that we know true teacher VA and thus can distinguish teachers accurately based on their quality. In reality, teacher VA is estimated. To account for this, we replace the test score gains in equations (7.2) and (7.5) with their sample analogs.[33] These sample analogs are calculated via Monte Carlo simulation, assuming that we estimate both $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty.

Under the PEB methodology, first we sample 40000 observations from $\mathcal{N}(0, \sigma_\alpha^2)$. Second, for each sample observation, we generate the noisy data $y_j = \alpha_j + \epsilon_j$ assuming $\epsilon_j \sim \mathcal{N}(0, \sigma_\epsilon^2/(k \cdot n))$, where n represents yearly class sizes (set at 20) and k represents the number of years of data for each teacher (set at 3). Third, we use the PEB estimator to obtain an estimated VA, δ_j^{PEB}, and calculate $\frac{1}{40000}\sum_j \alpha_j 1\{\delta_j^{PEB} > \hat{\Phi}_{\hat{\alpha}}^{-1}(q)\}$ as an estimator for $\widehat{TR}(q)$, the sample analog of equation (7.4). By the law of large numbers, this produces a consistent estimator for $\widehat{TR}(q)$. Analogously, we sample VA from the distribution F, use the NPEB method to obtain an estimator δ_j^{NPEB}, and calculate $\widehat{TR}(q)$ in a similar fashion.

Results based on these simulations are presented in Figure F.1. The policy gains fall when using estimated rather than true VA since some teachers with true VA below the fifth percentile are retained. The decrease in test score gains is relatively modest, however,

[33]For example, the marginal and total test score gains under $\alpha \sim F$ (equation (7.5)) become:

$$\widehat{MR}(q) = -\mathbb{E}_F[\alpha|\hat{\alpha} < \hat{F}_{\hat{\alpha}}^{-1}(q)]\,,$$
$$\widehat{TR}(q) = \mathbb{E}_F[\alpha \mathbb{1}\{\hat{\alpha} \geq \hat{F}_{\hat{\alpha}}^{-1}(q)\}]\,,$$

where $\hat{F}_{\hat{\alpha}}$ is the empirical CDF for the estimated $\hat{\alpha}$.

and is similar for both the PEB and NPEB methodologies. This is unsurprising since the methodologies do not substantially affect the ranking of teachers (as discussed in Section 6.1) and so using estimated rather than true VA should affect them in a similar manner.

Table 4(b) reports these policy gains when true teacher VA is unobserved. Results are very similar to the case when VA is observed by the policymaker: the PEB method overstates the policy gains by seven percent in North Carolina and by fully *twenty-seven* percent in the LAUSD data.

7.2 Policy Experiment II: Retention Policies

Next we consider policies that focus on high-VA teachers in the right tail of the distribution. Specifically, we examine policies that seek to retain these high-quality teachers, who might otherwise leave the profession. We assume that if a high-quality teacher is not retained, she is replaced with an average quality teacher (with a VA of zero). The total test score gains from a policy that retains teachers whose VA is greater than or equal to the $1-q$ percentile of the quality distribution are given by:

$$TR(q) = -\int_{-\infty}^{F^{-1}(1-q)} \alpha f(\alpha) d\alpha \,. \tag{7.7}$$

(Computational details are analogous to those for the 'teacher release' policy above.)

Tables 5(a) and 5(b) compare the policy gains for North Carolina and the LAUSD when true teacher quality is observable and unobservable (respectively) to the policymaker. Analogously, Figure F.2 plots these gains. Here, at the far right tail of the distribution (i.e., the top three percent of teachers), the PEB methodology *underestimates* the policy gains of targeting high-VA teachers by about five percent in North Carolina when true teacher VA is unobserved (see columns (3) and (4)). Intuitively, this underestimation comes from the fact that North Carolina's teacher VA distribution features a fat right tail and so the PEB method 'believes' there are fewer right-tail teachers than there actually are. As we move away from the extreme right tail of the distribution (to around the fifth percentile),

the tail becomes thinner and so the PEB methodology begins to overestimate the policy gains of targeting high VA teachers. A similar story holds in the LAUSD data, although the underestimation of the PEB estimator in the far right tail is less pronounced.

7.3 Extensions

With our data structure, according to which each teacher teaches one class of students each year – a feature common in many education data sets – our methodology can be extended to allow for *either* class-level shocks or drift in teacher quality. We note, however, that our data structure does not allow us to account for both simultaneously.[34] We focus on extending our model to allow for class-level shocks rather than drift since the amount of drift in both our data sets appears limited, consistent with prior research using these data sets.[35] That is, classroom shocks are quantitatively more important in our data.

Class-level Shocks: These are common shocks affecting everyone in a given classroom. The proverbial example involves a dog barking outside the classroom window on test day, lowering the test scores of all students in that class. Because teachers are unable to control the dog barking in this example, these class-level shocks should not be attributed to the teacher. Accounting for them acts to reduce dispersion in teacher VA estimates, as they subsume some of the class-year variation that was previously attributed to teachers. Since the variance of VA estimates falls once these shocks are incorporated, the policy gains from targeting the tails of the teacher VA distribution are likely to decrease as teachers in the tails are pulled closer to the mean. Our results below are consistent with this pattern.

Given that classrooms in our data are identified by unique teacher-year pairs, we rewrite

[34]To include both class-level shocks and drift, one would need data in which teachers taught multiple classrooms in each year, as in Chetty et al. (2014a). Those authors are able to allow for class-level shocks and drift for middle school teachers, observed to teach multiple classes each year.

[35]The small amount of drift (relative to Chetty et al. (2014a)) has been noted by Bacher-Hicks et al. (2014) for the LAUSD data and by Rothstein (2017) for the North Carolina data.

our model given by equation (2.2) as:

$$y_{ijt} = \alpha_j + \theta_{jt} + \epsilon_{ijt}, \quad i = 1, 2, \ldots, n_{jt}, \tag{7.8}$$

where y_{ijt} is student i's residual test score, α_j is teacher j's VA, and θ_{jt} represent the class-level shocks, which are independent of the student-level shocks, ϵ_{ijt}.

Assuming classroom shocks are distributed normally with variance σ_θ^2, it follows that the teacher-year specific sample mean can be modeled as $y_{jt} = \alpha_j + \nu_{jt}$ where $\nu_{jt} \sim N(0, \sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}})$. The teacher-specific fixed effect estimator y_j is then constructed as before (see equation (2.3)), except the weights h_{jt} now become $\left(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}}\right)^{-1}$. The NPEB estimator, parallel to the result in Theorem 2, can be expressed as follows:

Theorem 3 *Given the model $y_{jt} = \alpha_j + \nu_{jt}$, with $\alpha_j \sim F$, and $\nu_{jt} \sim \mathcal{N}(0, \sigma_\theta^2 + \sigma_\epsilon^2/n_{jt})$, the fixed effect estimator for α_j takes the form*

$$y_j = \sum_t h_{jt} y_{jt} / \sum_t h_{jt}$$

with $h_{jt} = \left(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}}\right)^{-1}$, and the estimator of α_j that minimizes the Bayes risk under $\mathcal{L}_2$ loss takes the form

$$\delta_j^{NPB} = y_j + \Big(\sum_t \left(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}}\right)^{-1} \Big)^{-1} \frac{\partial}{\partial y} \log g_j(y)|_{y=y_j} \,,$$

where $g_j(\cdot)$ is the marginal density of y_j.

Proof. The proof is very similar to that of Theorem 2 and follows from the fact that the fixed effects $\{y_j\}$ take the form $y_j = \alpha_j + \nu_j$, and $\nu_j \sim \mathcal{N}(0, (\sum_t h_{jt})^{-1})$. ■

The above theorem assumes a normal distribution for both the classroom and student-level shocks (i.e, $\theta_{jt} \sim \mathcal{N}(0, \sigma_\theta^2)$ and $\epsilon_{ijt} \sim \mathcal{N}(0, \sigma_\epsilon^2)$). Once again, we show that this is not necessary in Appendix A.2, although it is imposed for estimation purposes, with the

parameters σ_θ^2 and σ_ϵ^2 being estimated using maximum likelihood (as described in Appendix D.2).

Results: We incorporate classroom shocks in our estimation of the distribution of teacher quality. As expected, the addition of classroom shocks reduces the variance of our estimated distribution, with this reduction being more pronounced in the LAUSD data relative to the NC data. The reduction in variance is also larger when using the PEB methodology relative to our nonparametric method.

Revisiting the policy evaluation, Figure F.3 displays the policy gains from releasing the bottom q percent of teachers according to VA under both our nonparametric method and the PEB methodology when teacher quality is observed, then unobserved, respectively. As in the case without classroom shocks, the PEB methodology only overestimates policy gains by a small margin in North Carolina, but substantially overestimates them in the LAUSD. Relative to the model without classroom shocks, however, the overestimation in both data sets is reduced. This occurs since the addition of classroom shocks reduces the variance in the PEB methodology by more than in the NPEB methodology, and the q percent of teachers being released under a low-variance distribution will tend to be of higher quality relative to a high variance distribution.

Tables F.2(a) and F.2(b) report the estimated policy gains (along with their standard errors) when true teacher VA is observed and unobserved to the policymaker, respectively. When teacher VA is unobserved by the policymaker, the PEB method does *not* overestimate policy gains in North Carolina under our benchmark policy that releases bottom-ventile teachers (made bold in Table F.2(b)). In the LAUSD, however, policy gains are still overestimated by over *sixteen* percent.

Similarly, Tables F.3(a) and F.3(b) report the policy gains for teacher retention under the classroom shocks model. Specifically, the PEB method underestimates policy gains (as for the case without classroom shocks), with the degree of underestimation being higher than in the model without classroom shocks (as the decreased variance in the PEB relative to NPEB

methodology exacerbates the underestimation). Here, the PEB methodology underestimates the policy gains of retention policies targeting top-ventile teachers by around 15-20 percent (with this underestimation being more pronounced in the North Carolina data).

8 Conclusion

In this paper, we have proposed a new approach to estimating teacher VA that relaxes the normality assumption embedded in the popular parametric Empirical Bayes method. Our nonparametric Empirical Bayes approach is appealing in that it allows the underlying distribution of teacher quality to be estimated directly, and in a computationally feasible way using large data sets.

We applied the methodology to two separate administrative data sets in education, showing that the estimated teacher VA distributions differed from each other and departed from normality. We then explored the implications of these departures from normality in a range of policy evaluations, showing that the benefits of teacher lay-off policies may be overstated to a large degree (at least, in one of the two settings) and that the benefits of retention policies may be overstated (in the other setting).

The nonparametric approach to estimation has broader applicability to other areas of education research, where the underlying heterogeneity of students, teachers and schools is intrinsic. For example, looking beyond the current application, our methodology is well-suited to capturing dynamic policy-driven changes in underlying teacher quality distributions. Here, suppose that policymakers implemented a policy releasing teachers in the bottom of the teacher VA distribution *every* year. Under such a policy, the left tail of the teacher quality distribution would necessarily become truncated. When imposing a normality assumption in this case, 'fitting the data' would then require lowering the VA of teachers near the truncation point to 'create' a left tail, thereby underestimating the VA of teachers at the bottom of the distribution, in turn likely overestimating the gains of continuing the policy. Given that our method estimates such changes in the underlying teacher quality

distributions flexibly, it should provide sharper predictions regarding the continuing policy gains associated with implementing such dynamic reforms.

Our analysis has served to underline the notion that analytical convenience need no longer weigh on the side of assuming normality when applying empirical Bayes methods. The NPEB approach is relevant in a variety of other settings where parametric empirical Bayes methods have been used. These include, to date, the estimation of non-cognitive teacher effects (Jackson, 2018; Petek and Pope, 2018), school quality (Angrist et al., 2017; Bruhn, 2020), neighborhood effects (Chetty and Hendren, 2018), discrimination (Goncalves and Mello, 2018), physician effects (Fletcher et al., 2014), and hospital effects (Chandra et al., 2016; Hull, 2020). As large-scale panel data sets become more widely available in various fields, so the range of feasible applications using the nonparametric empirical Bayes approach is likely to increase. To that end, we have written, and are making available, code that will allow researchers to implement the NPEB method in other contexts.

References

Abadie, Alberto and Maximilian Kasy (2019), "Choosing among regularized estimators in empirical economics: The risk of machine learning." *Review of Economics and Statistics*, 101, 743–762.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters (2017), "Leveraging lotteries for school value-added: Testing and estimation." *Quarterly Journal of Economics*, 132, 871–919.

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger (2014), "Validating teacher effect estimates using changes in teacher assignments in Los Angeles." Working Paper 20657, National Bureau of Economic Research, URL `http://www.nber.org/papers/w20657`.

Bitler, Marianne, Sean Corcoran, Thurston Domina, and Emily Penner (2019), "Teacher effects on student achievement and height: A cautionary tale." Working Paper 26480, National Bureau of Economic Research, URL `http://www.nber.org/papers/w26480`.

Bonhomme, Stéphane and Martin Weidner (2019), "Posterior average effects." Working Paper CWP43/19, Centre for Microdata Methods and Practice, URL `https://www.cemmap.ac.uk/publication/id/14366`.

Brown, Lawrence D. (2008), "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies." *Annals of Applied Statistics*, 2, 113–152.

Brown, Lawrence D. and Eitan Greenshtein (2009), "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means." *Annals of Statistics*, 37, 1685–1704.

Bruhn, Jesse (2020), "The consequences of sorting for understanding school quality." URL `http://www.jessebruhn.com/research`. Unpublished.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson (2016), "Health care exceptionalism? Performance and allocation in the US health care sector." *American Economic Review*, 106, 2110–44.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014a), "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." *American Economic Review*, 104, 2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014b), "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review*, 104, 2633–79.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2017), "Measuring the impacts of teachers: Reply." *American Economic Review*, 107, 1685–1717.

Chetty, Raj and Nathaniel Hendren (2018), "The impacts of neighborhoods on intergenerational mobility II: County-level estimates." *Quarterly Journal of Economics*, 133, 1163–1228.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006), "Teacher-student matching and the assessment of teacher effectiveness." *Journal of Human Resources*, 41, 778–820.

Dicker, Lee H. and Sihai D. Zhao (2016), "High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference." *Biometrika*, 103, 21–34.

Efron, Bradley (2003), "Robbins, empirical Bayes and microarrays." *Annals of Statistics*, 31, 366–378.

Efron, Bradley (2010), *Large-scale Inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, Cambridge, UK.

Efron, Bradley (2011), "Tweedie's formula and selection bias." *Journal of American Statistical Association*, 106, 1602–1614.

Evdokimov, Kirill and Halbert White (2012), "Some extensions of a lemma of Kotlarski." *Econometric Theory*, 28, 925–932.

Fan, Jianqing (1991), "On the optimal rates of convergence for nonparametric deconvolution problems." *Annals of Statistics*, 19, 1257–1272.

Fletcher, Jason M., Leora I. Horwitz, and Elizabeth Bradley (2014), "Estimating the value added of attending physicians on patient outcomes." Working Paper 20534, National Bureau of Economic Research, URL `http://www.nber.org/papers/w20534`.

Goldhaber, Dan and Richard Startz (2017), "On the distribution of worker productivity: The case of teacher effectiveness and student achievement." *Statistics and Public Policy*, 4, 1–12.

Goncalves, Felipe and Steven Mello (2018), "A few bad apples? Racial bias in policing." URL `https://static1.squarespace.com/static/58d9a8d71e5b6c72dc2a90f1/t/5cfe39c1db1f980001595d4d/1560164805693/GoncalvesMello.pdf`. Unpublished.

Gu, Jiaying and Roger Koenker (2017a), "Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data." *Journal of Applied Econometrics*, 32, 575–599.

Gu, Jiaying and Roger Koenker (2017b), "Unobserved heterogeneity in income dynamics: An empirical Bayes perspective." *Journal of Business & Economic Statistics*, 35, 1–16.

Gu, Jiaying, Roger Koenker, and Stanislav Volgushev (2018), "Testing for homogeneity in mixture models." *Econometric Theory*, 34, 850 – 895.

Gu, Jiaying and Shu Shen (2017), "Oracle and adaptive false discovery rate controlling methods for one-sided testing: Theory and application in treatment effect evaluation." *Econometrics Journal*, 21, 11–35.

Guarino, Cassandra M., Michelle Maxfield, Mark D. Reckase, Paul N. Thompson, and Jeffrey M. Wooldridge (2015), "An evaluation of empirical Bayes's estimation of value-added

teacher performance measures." *Journal of Educational and Behavioral Statistics*, 40, 190–222.

Hanushek, Eric A. (2009), "Teacher deselection." In *Creating a New Teaching Profession* (Dan Goldhaber and Jane Hannaway, eds.), 165–180, Urban Institute Press, Washington, DC.

Hanushek, Eric A. (2011), "The economic value of higher teacher quality." *Economics of Education Review*, 30, 466–479.

Harris, Douglas N. and Tim R. Sass (2014), "Skills, productivity and the evaluation of teacher performance." *Economics of Education Review*, 40, 183–204.

Heckman, James and Burton Singer (1984), "A method for minimizing the impact of distributional assumptions in econometric models for duration data." *Econometrica*, 52, 271–320.

Hull, Peter (2020), "Estimating hospital quality with quasi-experimental data." URL `https://www.google.com/url?q=https%3A%2F%2Fwww.dropbox.com%2Fs%2Fhb54rrz3vte8gij%2FRAM_012020.pdf%3Fraw%3D1&sa=D&sntz=1&usg=AFQjCNE-ap7RlsV8PsFpJ64ekwD2HmqIjQ`. Unpublished.

Jackson, C. Kirabo (2018), "What do test scores miss? The importance of teacher effects on non-test score outcomes." *Journal of Political Economy*, 126, 2072–2107.

Jacob, Brian A. and Lars Lefgren (2008), "Can principals identify effective teachers? Evidence on subjective performance evaluation in education." *Journal of Labor Economics*, 26, 101–136.

James, W. and Charles Stein (1961), "Estimation with quadratic loss." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 361–379, University of California Press, Berkeley, Calif.

Jiang, Wenhua and Cun-Hui Zhang (2009), "General maximum likelihood empirical Bayes estimation of normal means." *Annals of Statistics*, 37, 1647–1684.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger (2008), "What does certification tell us about teacher effectiveness? Evidence from New York City." *Economics of Education Review*, 27, 615–631.

Kane, Thomas J. and Douglas O. Staiger (2008), "Estimating teacher impacts on student achievement: An experimental evaluation." Working Paper 14607, National Bureau of Economic Research, URL `http://www.nber.org/papers/w14607`.

Kiefer, Jack and Jacob Wolfowitz (1956), "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters." *Annals of Mathematical Statistics*, 27, 887–906.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff (2015), "Value-added modeling: A review." *Economics of Education Review*, 47, 180–195.

Koenker, Roger and Jiaying Gu (2017), "Rebayes: An R package for empirical Bayes mixture methods." *Journal of Statistical Software*, 82, 1–26.

Koenker, Roger and Ivan Mizera (2014), "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules." *Journal of the American Statistical Association*, 109, 674–685.

Kotlarski, Ignacy (1967), "On characterizing the gamma and the normal distribution." *Pacific Journal of Mathematics*, 20, 69–76.

Laird, Nan (1978), "Nonparametric maximum likelihood estimation of a mixing distribution." *Journal of the American Statistical Association*, 73, 805–811.

Laird, Nan M. and Thomas A. Louis (1987), "Empirical Bayes confidence intervals based on bootstrap samples." *Journal of American Statistical Association*, 82, 805–811.

Li, Tong and Quang Vuong (1998), "Nonparametric estimation of the measurement error model using multiple indicators." *Journal of Multivariate Analysis*, 65, 139–165.

McLachlan, G.J. (1987), "On bootstrapping likelihood ratio test statistics for the number of components in a normal mixture." *Journal of the Royal Statistical Society, Series C*, 36, 318–324.

Mehta, Nirav (2019), "Measuring quality for use in incentive schemes: The case of "shrinkage" estimators." *Quantitative Economics*, 10, 1537–1577.

Petek, Nathan and Nolan Pope (2018), "The multidimensional impact of teachers on students." URL `http://www.econweb.umd.edu/~pope/Nolan_Pope_JMP.pdf`. Unpublished.

Rao, B.L.S.P. (1992), *Identifiability in Stochastic Models: Charaterization of Probability Distributions*. Academic Press, United Kindom.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005), "Teachers, schools, and academic achievement." *Econometrica*, 73, 417–458.

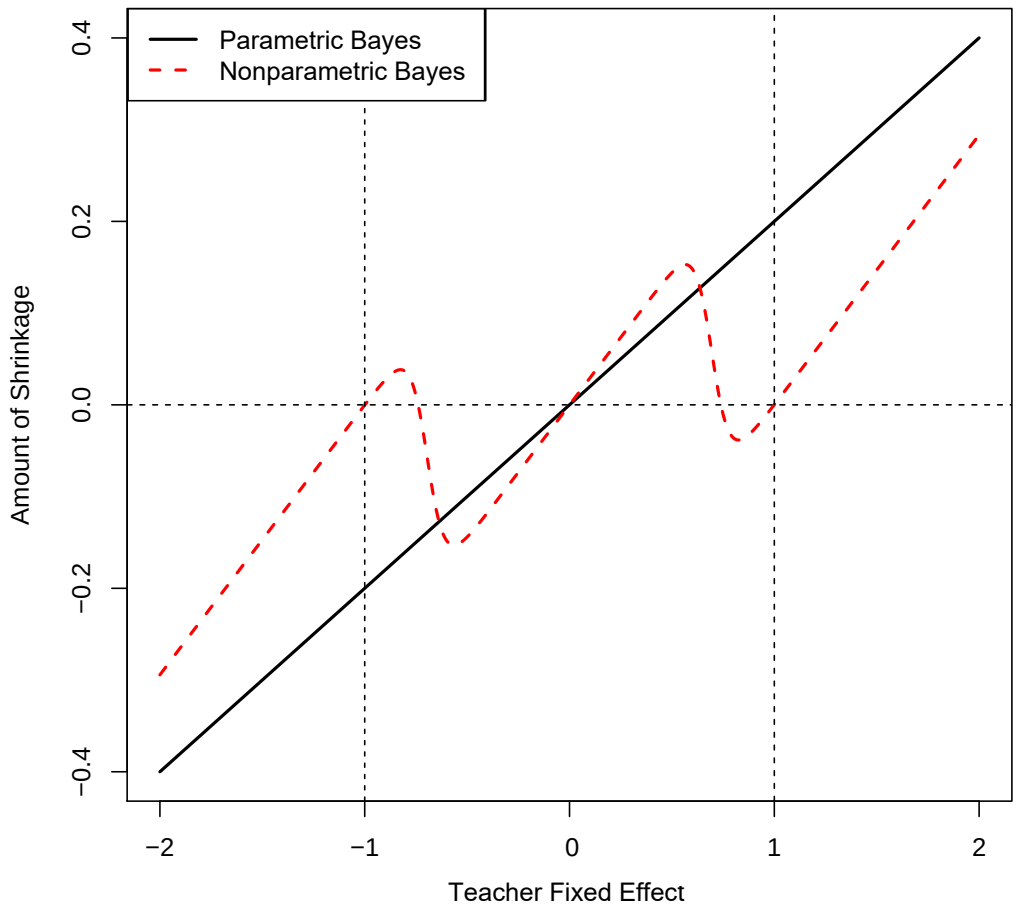
Robbins, Herbert (1956), "An empirical Bayes approach to statistics." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume I, University of California Press, Berkeley.

Rockoff, Jonah E. (2004), "The impact of individual teachers on student achievement: Evidence from panel data." *American Economic Review*, 94, 247–252.

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

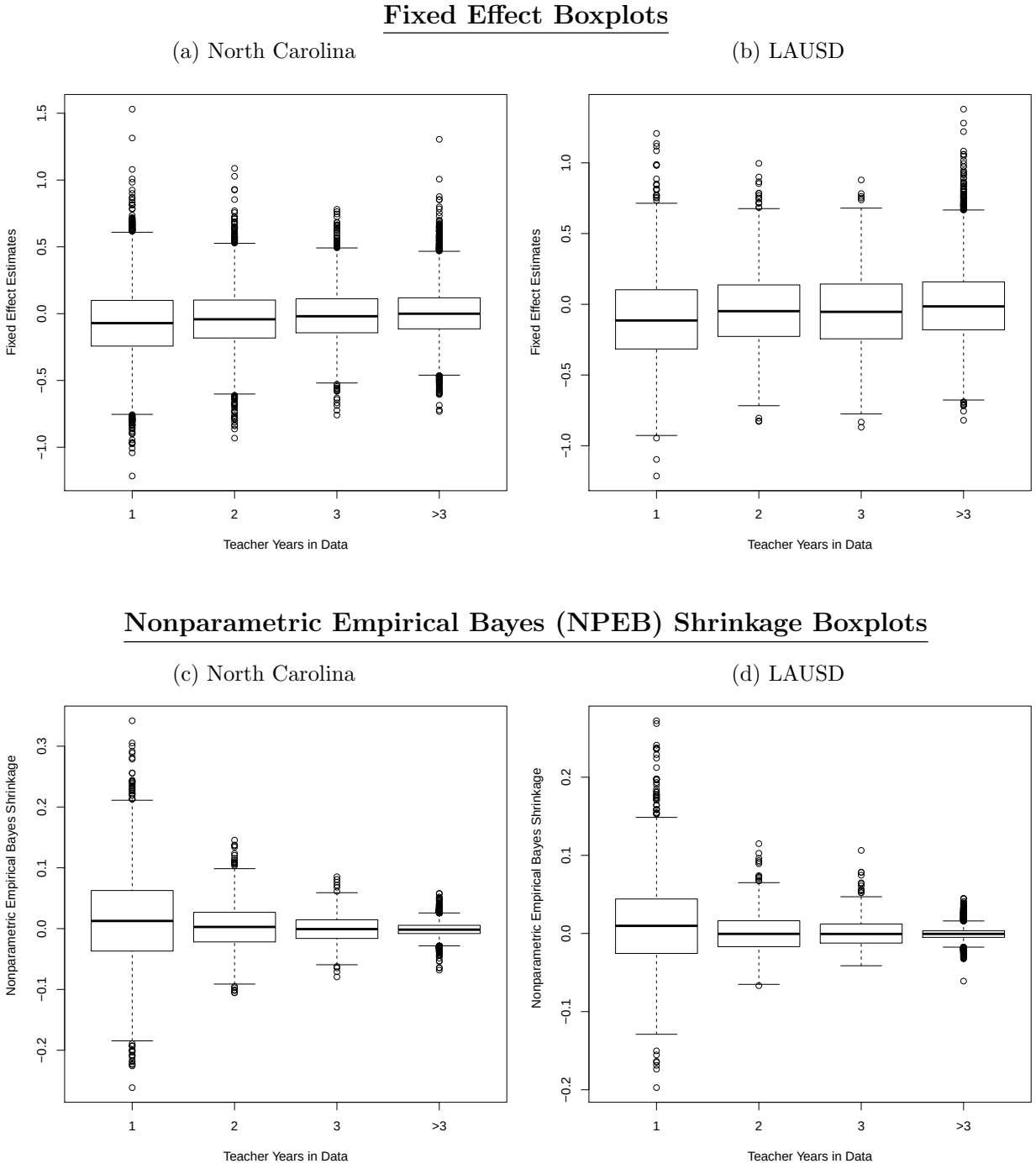
Saha, Sujayam and Adityanand Guntuboyina (forthcoming), "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising." *Annals of Statistics*.

Figure 1: Example of Shrinkage under Parametric and Nonparametric Empirical Bayes Estimators



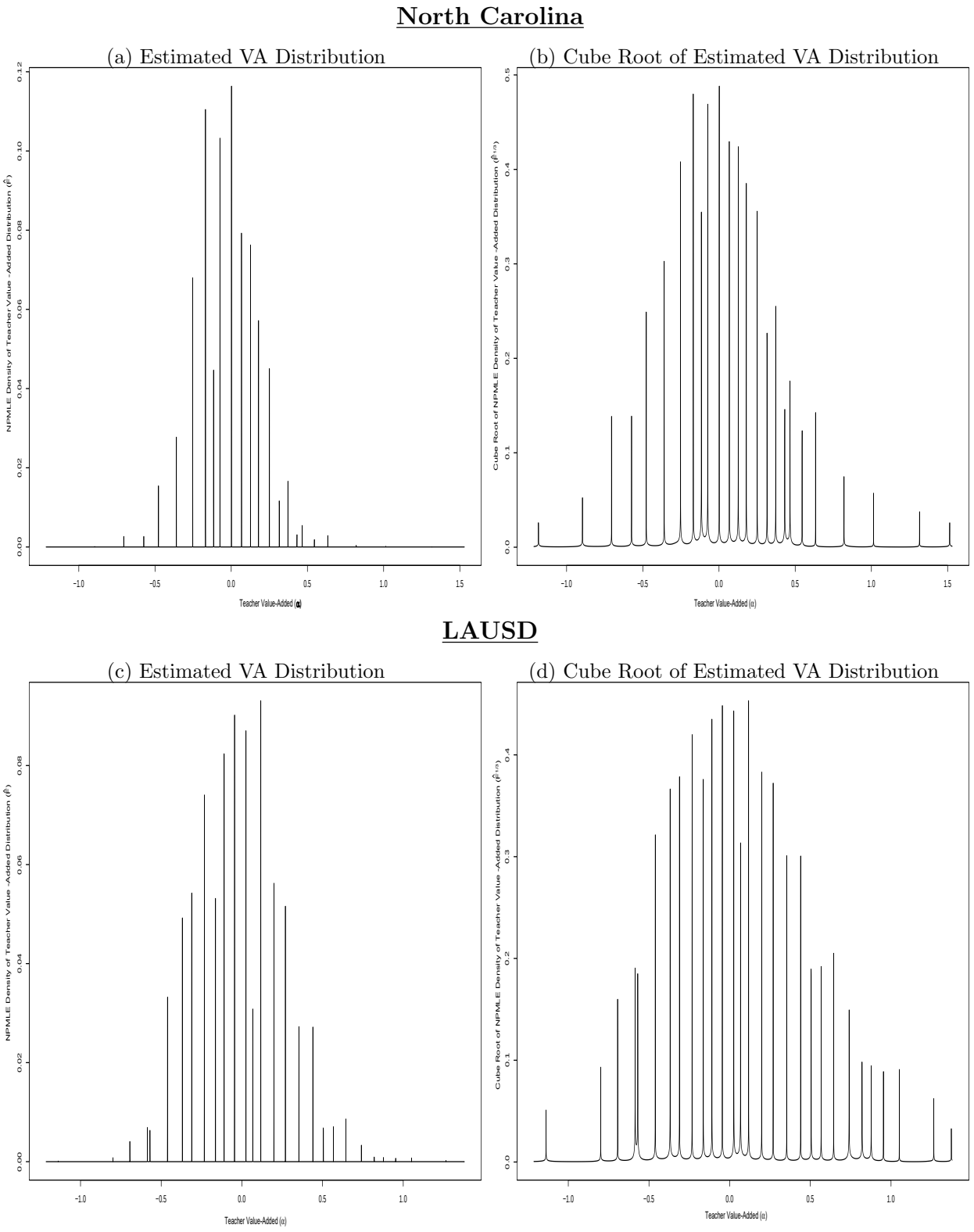
Notes: This figure plots the amount of Bayesian 'shrinkage' as a function of the fixed effect estimates for the PB and NPEB estimators respectively. The shrinkage rule for the PB estimator is given by $\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2/\sum_t n_{jt}}$ from equation (2.5); for the NPB estimator, it is given by the second term in equation (2.7). It does so for the mixed normal distribution specified in equation (2.8), given by $\alpha_j \sim 0.98\mathcal{N}(0, 0.03) + 0.01\mathcal{N}(-1, 0.03) + 0.01\mathcal{N}(1, 0.03)$. The total class size for each teacher is set at twenty and σ_ϵ^2 is set at 0.25. Fixed effects take values in the range $[-2, 2]$. The horizontal dashed line represents no shrinkage being applied, while the vertical dashed lines represent the mass points in the distribution at -1 and $+1$.

Figure 2: Boxplots of Fixed Effects and Shrinkage under Nonparametric Empirical Bayes



Notes: Figures 2(a) and 2(b) give the raw fixed effect estimates by the number of times a teacher appears in the data. Specifically, each panel displays a boxplot of fixed effect estimates for teachers who appear once, twice, three times or more than three times in our North Carolina and LAUSD datasets, respectively. Figures 2(c) and 2(d) then show boxplots of the amount of shrinkage applied by our NPEB estimator to teachers who appear once, twice, three times or more than three times in our North Carolina and LAUSD datasets, respectively. (Boxplots use the box to indicate the interquartile range between the first and third quartile and use whiskers to indicate the first (respectively, third) quartile minus (plus) the interquartile range multiplied by 1.5. Outliers beyond this range are shown with dots.)

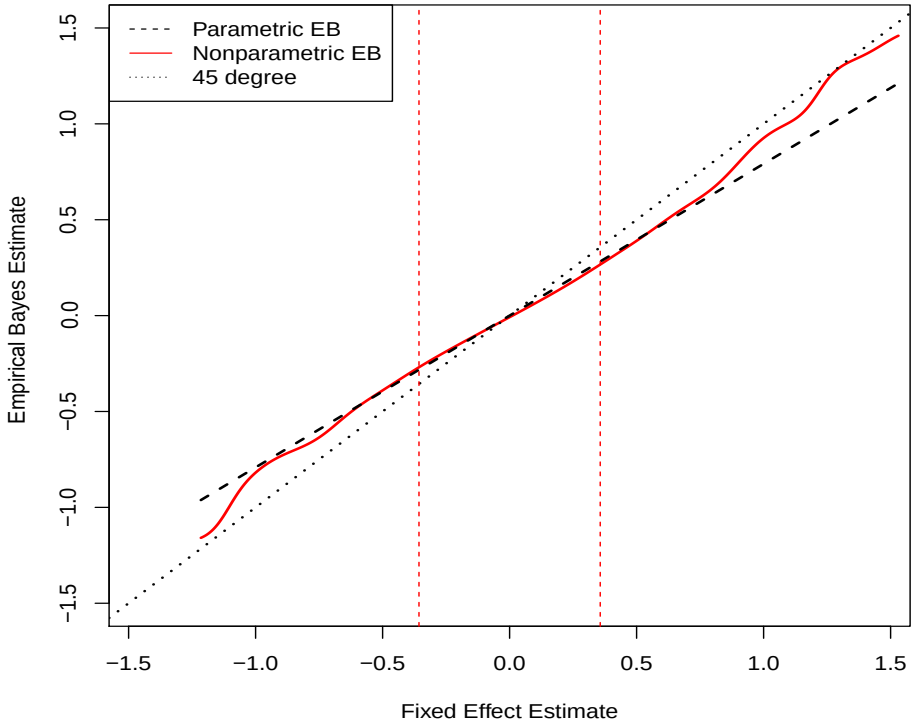
Figure 3: Estimated Teacher Quality Distributions using Nonparametric Empirical Bayes



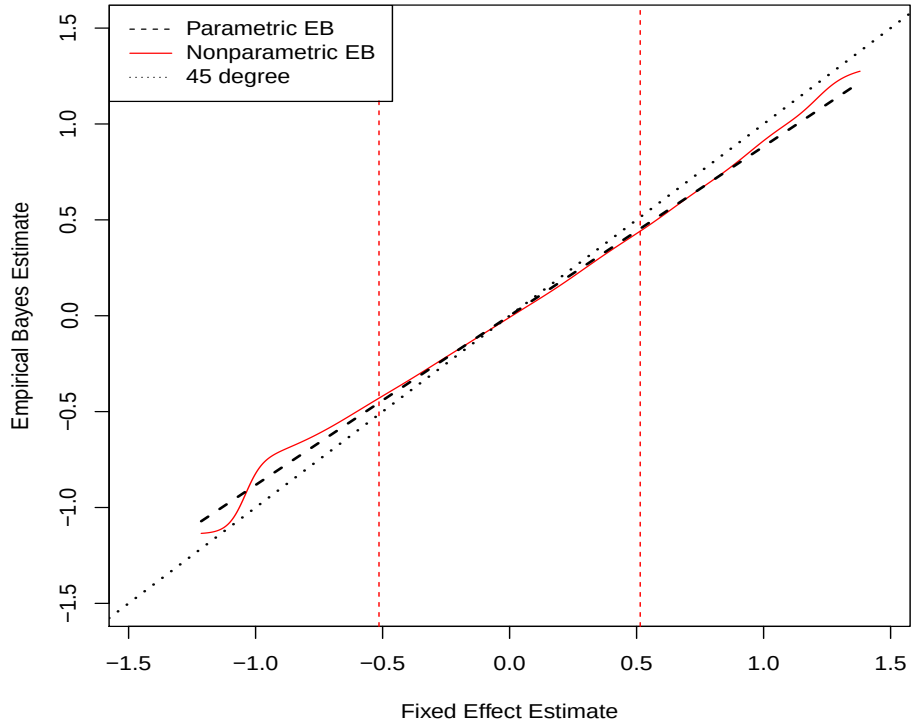
Notes: Figures 3(a) and 3(c) display the estimated distribution of teacher quality (VA), $\hat{F}$, for North Carolina and the LAUSD, respectively. These distributions are estimated nonparametrically using equation (3.1). In order to better see tail behavior, Figures 3(b) and 3(d) take a cube root of the estimated VA distribution to boost the tails of the distribution for North Carolina and the LAUSD, respectively.

Figure 4: Empirical Bayes 'Shrinkage' for Fixed Class Size
(class size = 20)

(a) North Carolina

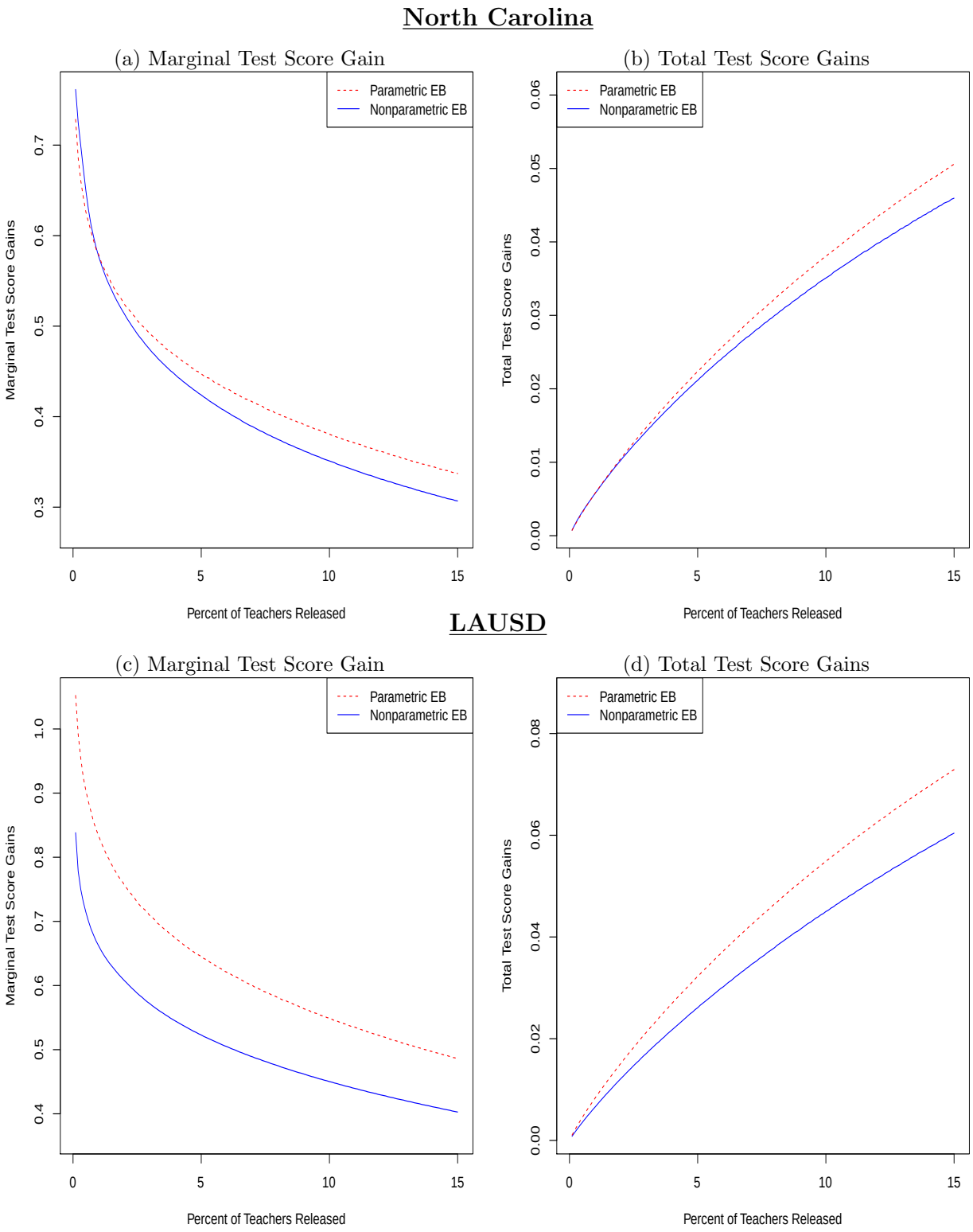


(b) LAUSD



Notes: Figures 4(a) and 4(b) show fixed effect estimates relative to the Bayes estimates for both the parametric (PEB) and nonparametric empirical Bayes (NPEB) methodologies. The dotted line represents the 45 degree line and indicates where the fixed effect and empirical Bayes estimates agree. Since the amount of Bayes 'shrinkage' applied depends on the total number of students taught by the teacher, we display the rule for a representative teacher who has taught a total class size of twenty students. The vertical dashed lines represent the 5^{th} and 95^{th} percentiles of teacher VA estimates according to the fixed effect estimates to delineate the tails of the value-added distribution.

Figure 5: Test Scores Gains from Replacing the Bottom q Percentile of Teachers



Notes: Figures 5(a) and 5(b) show the marginal and total test score gains of a policy that releases the bottom q% of teachers in North Carolina, while figures 5(c) and 5(d) do the same for the LAUSD. The dotted lines indicate the policy gains expected under the PEB methodology assuming that $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, where σ_α^2 is estimated from the data. The solid lines denote the policy gains expected under our NPEB methodology where we allow $\alpha \sim F$, with F being estimated directly from the data using equation (3.1). Policy gains reported here assume that policymakers observe true underlying value-added; Figure F.1 presents the estimated gains if value-added is estimated rather than observed.

Table 1(a): Simulation – *True Distribution is Normal*

	Homogeneous Class Sizes (Class Size of 20)				Heterogeneous Class Sizes (Class Size 20-40)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Mean Squared Error	10.81	10.87	10.81	12.50	11.03	11.07	11.03	12.84

Table 1(b): Simulation – *True Distribution is Mixed Normal*

	Homogeneous Class Sizes (Class Size of 20)				Heterogeneous Class Sizes (Class Size 20-40)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Mean Squared Error	9.08	9.14	10.82	12.51	7.14	7.18	8.29	9.36

Table 1(c): Simulation – *True Distribution is Chi-Squared*

	Homogeneous Class Sizes (Class Size of 20)				Heterogeneous Class Sizes (Class Size 20-40)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Mean Squared Error	7.44	7.45	10.82	12.51	5.79	5.79	8.30	9.37

Notes: The panels in this table report simulation results comparing (across the columns) the performance of the three candidate estimators against an infeasible benchmark on the basis of mean squared error. The infeasible benchmark is the optimal estimator when the true distribution is known to the econometrician (although unknown in practice). The three candidate estimators are: the nonparametric empirical Bayes (NPEB) estimator, which estimates the underlying distribution nonparametrically; the parametric empirical Bayes (PEB) estimator, which assumes that the underlying distribution is normal; and the fixed effect (FE) estimator, which applies no empirical Bayes shrinkage. In Table 1(a), teacher VA is normally distributed with mean zero and variance 0.05 (i.e., $F \sim \mathcal{N}(0, 0.05)$). In Table 1(b), true teacher quality follows $F \sim 0.98\mathcal{N}(0, 0.03) + 0.01\mathcal{N}(-1, 0.03) + 0.01\mathcal{N}(1, 0.03)$ (as in the example given by equation (2.8)). The normal and mixed normal distributions have the same mean and variance for comparability. In Table 1(c), teacher value-added follows $F \sim \chi^2_1$. The simulations average results from 500 repetitions with 10,000 individual teachers setting $\sigma^2_\epsilon = 0.25$. Results are reported on the left side of each panel ((a)-(c)) for homogeneous class sizes (where every teacher has a class size of twenty) and on the right side, for heterogeneous class sizes (where class sizes are drawn randomly from the set $\{20, 40\}$ with equal probability).

Table 2: Summary Statistics

	North Carolina		LAUSD	
	Full Sample[1]	Value-Added Sample	Full Sample[2]	Value-Added Sample
	(1)	(2)	(3)	(4)
Mean of Student Characteristics				
Math Score (σ)	0.00	0.05	0.00	0.07
Reading Score (σ)	0.00	0.03	0.00	0.06
Lagged Math Score (σ)	0.01	0.03	0.03	0.08
Lagged Reading Score (σ)	0.01	0.03	0.03	0.07
% White	57.8	60.1	9.3	9.1
% Black	28.8	27.9	9.9	8.6
% Hispanic	7.4	6.5	74.0	75.5
% Asian	2.0	1.9	4.3	4.4
% Free or Reduced Price Lunch[3]	46.3	44.6	77.9	78.2
% English Learners	4.3	3.5	28.0	28.9
% Repeating Grade	1.5	1.5	1.5	0.4
Parental Education:[4]				
% High School Dropout	11.5	10.6	34.5	34.4
% High School Graduate	47.31	47.0	27.6	27.8
% College Graduate	25.4	25.9	20.1	20.0
Teacher Experience:[5]				
0-2 Years of Experience	18.6	18.8	4.9	4.8
3-5 Years of Experience	15.3	15.6	10.5	10.3
# of Students	1,847,615	1,386,555	810,753	664,044
# of Teachers	76,503	35,053	15,267	11,078
Observations[6] (student-year)	4,457,812	2,680,027	1,707,459	1,280,569

Notes:

[1] North Carolina data coverage: grades 4-5 from 1996-97 through 2010-11 and grade 3 from 1996-97 through 2009-10. The difference in sample sizes between columns (1) and (2) is because we drop 1.37 million student-year observations who we cannot match to their classroom teacher (see Appendix B for more detail).

[2] Los Angeles Unified School District (LAUSD) data coverage: grades 4-5 from 2003-04 through 2012-13 and 2015-16 through 2016-17 school years and third grade from 2003-04 through 2012-13.

[3] For North Carolina this variable is missing for school years 1996-97 through 1997-98.

[4] The omitted category is 'Some College,' and 'College Graduate' also incorporates those with graduate school degrees. For North Carolina, parental education data are missing after the 2005-06 school year, while thirty percent of observations in the LAUSD are missing parental education data or have parental education recorded as "Decline to Answer."

[5] The omitted category is 'Greater than 5 Years of Experience.' For the full sample, teacher experience data are missing for about twenty and fifteen percent of observations for North Carolina and LAUSD, respectively.

[6] Data are missing for some observations. For North Carolina (full sample), test scores are missing for three percent of observations, lagged test scores for twelve percent, with most other other demographic variables missing for about one percent of observations. For the LAUSD (full sample), lagged test scores are missing for about six percent of observations with data coverage for all other variables near one hundred percent.

Table 3(a): Predicted Performance of Nonparametric Empirical Bayes (NPEB), Parametric Empirical Bayes (PEB) and Fixed Effects – North Carolina Data

# of Prior Years Used	NMSE			TMSE		
	NPEB	PEB	Fixed Effects	NPEB	PEB	Fixed Effects
$t = 1$	1.0096	1.0098	1.2236	0.0378	0.0378	0.0486
$t = 2$	0.8613	0.8715	0.9397	0.0304	0.0309	0.0343
$t = 3$	0.7784	0.7872	0.8229	0.0261	0.0265	0.0283
$t = 4$	0.7704	0.7767	0.7998	0.0259	0.0262	0.0273
$t = 5$	0.7651	0.7708	0.7857	0.0257	0.0260	0.0267
$t = 6$	0.7532	0.7573	0.7687	0.0250	0.0252	0.0258
$t = 7$	0.7255	0.7294	0.7372	0.0240	0.0242	0.0245
$t = 8$	0.7123	0.7161	0.7240	0.0236	0.0238	0.0242

Table 3(b): Predicted Performance of Nonparametric Empirical Bayes (NPEB), Parametric Empirical Bayes (PEB) and Fixed Effects – LAUSD Data

# of Prior Years Used	NMSE			TMSE		
	NPEB	PEB	Fixed Effects	NPEB	PEB	Fixed Effects
$t = 1$	1.6205	1.6180	1.7975	0.0631	0.0630	0.0714
$t = 2$	1.4030	1.4084	1.4634	0.0526	0.0529	0.0554
$t = 3$	1.3865	1.3902	1.4138	0.0510	0.0512	0.0523
$t = 4$	1.3929	1.3970	1.4138	0.0513	0.0514	0.0522
$t = 5$	1.3852	1.3869	1.3978	0.0505	0.0506	0.0511
$t = 6$	1.4984	1.4999	1.5066	0.0538	0.0538	0.0541
$t = 7$	1.5209	1.5221	1.5306	0.0539	0.0539	0.0543
$t = 8$	1.4249	1.4254	1.4331	0.0496	0.0496	0.0499

Notes: Smaller values indicate better prediction performance, with NMSE (see equation (6.1)) and TMSE (equation (6.2)) representing normalized mean squared error and total mean squared error, respectively. Tables 3(a) and 3(b) report out-of-sample prediction errors in the North Carolina and LAUSD datasets for three different estimators: nonparametric empirical Bayes (NPEB), parametric empirical Bayes (PEB) and fixed effects. To deal with the variation in class size that teachers face across years, we use NMSE and TMSE as proposed by Brown (2008). The prediction performance is calculated by calculating the squared error distance (plus an adjustment term for class size) between the true outcome of teacher j in period $t+1$ and the outcome predicted for teacher j utilizing all past information relating to her teaching performance from period t minus the number of prior years used up until the t-th period. For each row, we subset the data so that each teacher is observed for at least $t+1$ periods.

Table 4(a): Test Scores Gains from Releasing Bottom q% of Teachers
True VA Observed

% Teachers Released (q)	North Carolina Data		LAUSD Data	
	Test Gain under F (NPEB) (1)	Test Gain under Normal (EB) (2)	Test Gain under F (NPEB) (3)	Test Gain under Normal (EB) (4)
1	0.0058	0.0058	0.0066	0.0083
	(0.0001)	(0.0000)	(0.0002)	(0.0001)
3	0.0142	0.0147	0.0171	0.0213
	(0.0002)	(0.0001)	(0.0003)	(0.0002)
5	**0.0212**	**0.0224**	**0.0261**	**0.0323**
	(0.0002)	(0.0001)	(0.0004)	(0.0002)
7	0.0272	0.0291	0.0342	0.0420
	(0.0003)	(0.0001)	(0.0004)	(0.0003)
9	0.0326	0.0352	0.0415	0.0508
	(0.0003)	(0.0002)	(0.0005)	(0.0004)

Table 4(b): Test Scores Gains from Releasing Bottom q% of Teachers
True VA Unobserved

% Teachers Released (q)	North Carolina Data		LAUSD Data	
	Test Gain under F (NPEB) (1)	Test Gain under Normal (EB) (2)	Test Gain under F (NPEB) (3)	Test Gain under Normal (EB) (4)
1	0.0056	0.0059	0.0064	0.0086
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
3	0.0137	0.0145	0.0166	0.0213
	(0.0003)	(0.0002)	(0.0003)	(0.0002)
5	**0.0203**	**0.0219**	**0.0254**	**0.0322**
	(0.0003)	(0.0003)	(0.0003)	(0.0003)
7	0.0261	0.0284	0.0332	0.0417
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
9	0.0312	0.0342	0.0405	0.0503
	(0.0005)	(0.0005)	(0.0005)	(0.0005)

Notes: Table 4(a) displays the estimated gains in mathematics scores in terms of student-level standard deviations of a policy that releases the bottom q% of teachers and replaces them with mean quality teachers when true teacher quality is observed by the policymaker. Reported policy gains are those also plotted in Figure 5. 'Test score gain under F' reports the test score gain of the policy when teacher quality is distributed according the distribution F – nonparametrically estimated using equation (3.1) – and applying the NPEB estimator to calculate value-added. 'Test score gain under normal' reports the test score gain when teacher quality is normally distributed and the PEB estimator is used to calculate teacher value-added. Table 4(b) repeats the exercise when the true teacher quality is unobserved to the policymaker and so teacher releases are based on estimated (rather than true) value-added. These gains are the same as those in Figure F.1 and are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. The bolded line indicates the widely-analyzed release bottom five percent teachers policy. Standard errors are calculated using bootstrap as described in Appendix C.

Table 5(a): Test Scores Gains of Policy Retaining Top $1-q$ Percentile Teachers
True VA Observed

% Teachers Retained $(1-q)$	North Carolina Data: Test Gain under F (NPEB) (1)	North Carolina Data: Test Gain under Normal (EB) (2)	LAUSD Data: Test Gain under F (NPEB) (3)	LAUSD Data: Test Gain under Normal (EB) (4)
1	0.0061	0.0058	0.0084	0.0083
	(0.0001)	(0.0000)	(0.0002)	(0.0001)
3	0.0147	0.0148	0.0206	0.0213
	(0.0002)	(0.0001)	(0.0003)	(0.0002)
5	0.0219	0.0224	0.0307	0.0323
	(0.0002)	(0.0001)	(0.0005)	(0.0002)
7	0.0281	0.0291	0.0394	0.0420
	(0.0003)	(0.0001)	(0.0005)	(0.0003)
9	0.0335	0.0352	0.0473	0.0508
	(0.0003)	(0.0002)	(0.0006)	(0.0004)

Table 5(b): Test Scores Gains of Policy Retaining Top $1-q$ Percentile Teachers
True Value-Added Unobserved

% Teachers Retained $(1-q)$	North Carolina Data: Test Gain under F (NPEB) (1)	North Carolina Data: Test Gain under Normal (EB) (2)	LAUSD Data: Test Gain under F (NPEB) (3)	LAUSD Data: Test Gain under Normal (EB) (4)
1	0.0060	0.0052	0.0084	0.0076
	(0.0001)	(0.0001)	(0.0002)	(0.0001)
3	0.0143	0.0137	0.0204	0.0202
	(0.0003)	(0.0002)	(0.0004)	(0.0002)
5	0.0212	0.0210	0.0302	0.0309
	(0.0003)	(0.0003)	(0.0005)	(0.0003)
7	0.0271	0.0274	0.0388	0.0404
	(0.0004)	(0.0004)	(0.0005)	(0.0004)
9	0.0324	0.0333	0.0465	0.0490
	(0.0005)	(0.0005)	(0.0006)	(0.0004)

Notes: Tables 5(a) displays the estimated gains in mathematics scores in terms of student level standard deviations of a policy that retains the top q% of teachers rather than having them leave teaching and be replace by a mean quality teacher when true teacher quality is observed by the policymaker. 'Test score gain under F' reports the test score gain of the policy when teacher quality is distributed according the distribution F – nonparametrically estimated using equation (3.1) – and applying the NPEB estimator to calculate value-added. 'Test score gain under normal' reports the test score gain when teacher quality is normally distributed and the PEB estimator is used to calculate teacher value-added. Table 5(b) repeats the exercise when the true teacher quality is unobserved to the policymaker and so teacher retentions are based on estimated (rather than true) value-added. These gains are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. Reported policy gains in both tables are identical to those shown in Figure F.2. Standard errors are calculated using bootstrap as described in Appendix C.

A General Deconvolution Proof with Panel Data

This appendix sets out the general deconvolution proof for the teacher VA model, first without, then with, classroom shocks.

A.1 Teacher VA model without classroom shocks

Assumption 1 *$Y_1 = \alpha + \epsilon_1$ and $Y_2 = \alpha + \epsilon_2$ where Y_1 and Y_2 are random variables with joint pdf $f(\cdot, \cdot)$, α is a random variable with pdf $g(\cdot)$, and ϵ_1 and ϵ_2 are random variables from the same pdf $h(\cdot)$ with mean zero.*

Assumption 2 *α, ϵ_1, and ϵ_2 are mutually independent.*

Assumption 3 *The characteristic functions $\phi_\alpha(\cdot)$ and $\phi_\epsilon(\cdot)$ of α and ϵ are nonvanishing everywhere.*

Lemma 1 (Kotlarski (1967)) *Under Assumption 1-3, the pdf's of α and ϵ are uniquely determined by the joint distribution of (Y_1, Y_2). In particular, let $\psi(u, v)$ be the characteristic function of the random vector (Y_1, Y_2), $\phi_\alpha(t)$ the characteristic function of α, and $\phi_\epsilon(t)$ the characteristic function of ϵ, then*

$$\phi_\alpha(t) = \exp \int_0^t \frac{\partial \psi(0, v)/\partial u}{\psi(0, v)} dv$$
$$\phi_\epsilon(t) = \frac{\psi(t, 0)}{\phi_\alpha(t)} = \frac{\psi(0, t)}{\phi_\alpha(t)}.$$

Proof. Using equation (2.64) in Rao (1992), we have

$$\log \phi_\alpha(t) = \mathtt{i}\mathbb{E}[\alpha]t + \int_0^t \frac{\partial}{\partial u} \Big(\log \frac{\psi(u, v)}{\psi(u, 0)\psi(0, v)} \Big)_{u=0} dv.$$

where $\mathtt{i}$ is the imaginary root. Using the fact that

$$\frac{\partial}{\partial u}\Big(\log\frac{\psi(u,v)}{\psi(u,0)\psi(0,v)}\Big)_{u=0} = \frac{\partial\psi(0,v)/\partial u}{\psi(0,v)} - \frac{\partial\psi(0,0)/\partial u}{\psi(0,0)}$$

and that $\frac{\partial\psi(0,0)/\partial u}{\psi(0,0)} = \mathtt{i}\mathbb{E}(Y_1)$, we have

$$\log\phi_\alpha(t) = \mathtt{i}\mathbb{E}[\alpha]t + \int_0^t \frac{\partial\psi(0,v)/\partial u}{\psi(0,v)}dv - \mathtt{i}\mathbb{E}(Y_1)t = \int_0^t \frac{\partial\psi(0,v)/\partial u}{\psi(0,v)}dv,$$

where the second equality holds because ϵ_1 has mean zero under Assumption 1.

Additionally, under Assumptions 1- 3, we have

$$\psi(u,v) = \phi_\alpha(u+v)\phi_\epsilon(u)\phi_\epsilon(v).$$

Let $u=0$, then $\phi_\epsilon(v) = \psi(0,v)/\phi_\alpha(v)$; and letting $v=0$, then $\phi_\epsilon(v) = \psi(u,0)/\phi_\alpha(u)$. ■

We note that Assumption 1 can be relaxed further to allow ϵ_1 and ϵ_2 to have different pdf's. (Recently, a relaxation of Assumption 3 is discussed in Evdokimov and White (2012).) Li and Vuong (1998) proposed a nonparametric plug-in estimator for $\phi_\alpha(t)$ and $\phi_\epsilon(t)$ through the nonparametric estimator for $\psi(\cdot,\cdot)$, based on J independent observations $\{(y_{1j}, y_{2j})\}_{j=1,\ldots,J}$ of (Y_1, Y_2), defined as

$$\hat{\psi}(u,v) = \frac{1}{J}\sum_{j=1}^{J}\exp(\mathtt{i}uy_{1j} + \mathtt{i}vy_{2j})$$

where, again, $\mathtt{i}$ is the imaginary unit. We then apply the inverse Fourier transform on $\phi_\alpha(t)$ and $\phi_\epsilon(t)$, yielding the density functions of α and ϵ.

Corollary 4 *Consider the general repeated measurement model,*

$$Y_{js} = \alpha_j + \epsilon_{js}, \quad j = 1, 2, \ldots, J \text{ and } s = 1, 2, \ldots, n_j,$$

where α is a random variable with pdf $g(\cdot)$ and the ϵ_s (with $s = 1, 2, \ldots, n_j$) are random variables from the same pdf $h(\cdot)$ with mean zero. If $n_j \geq 2$, α and ϵ_s are mutually independent, and the characteristic functions $\phi_\alpha(\cdot)$ and $\phi_\epsilon(\cdot)$ are nonvanishing everywhere, then the pdf's of α and ϵ are nonparametrically identified.

The above corollary applies for the teacher value-added model without classroom shocks, with j indexing teachers and n_j being the total number of students taught by teacher j. We can naturally construct the nonparametric estimator for $\psi(\cdot, \cdot)$ as

$$\hat{\psi}(u, v) = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{n_j(n_j - 1)} \sum_{1 \leq s_1 \neq s_2 \leq n_j} \exp(\mathtt{i}uy_{js_1} + \mathtt{i}vy_{js_2}).$$

A.2 Teacher VA model with classroom shocks

The above reasoning can be extended to the case where we allow for classroom shocks. To that end, we make three further assumptions:

Assumption 4 *$Y_{11} = \alpha + \theta_1 + \epsilon_{11}$, $Y_{21} = \alpha + \theta_1 + \epsilon_{21}$, $Y_{12} = \alpha + \theta_2 + \epsilon_{12}$, and $Y_{22} = \alpha + \theta_2 + \epsilon_{22}$ where Y_{11}, Y_{21}, Y_{12}, and Y_{22} are random variables with joint pdf $f(\cdot, \cdot, \cdot, \cdot)$, α is a random variable with pdf $g(\cdot)$, θ_1 and θ_2 are random variables from the same pdf $q(\cdot)$ with mean zero and ϵ_{11}, ϵ_{12}, ϵ_{21}, and ϵ_{22} are random variables from the same pdf $h(\cdot)$ with mean zero.*

Assumption 5 *α, θ_1, θ_2, ϵ_{11}, ϵ_{12}, ϵ_{21}, and ϵ_{22} are mutually independent.*

Assumption 6 *The characteristic functions $\phi_\alpha(\cdot)$, $\phi_\theta(\cdot)$ and $\phi_\epsilon(\cdot)$ of α, θ and ϵ are nonvanishing everywhere.*

Lemma 2 *Under Assumptions 4 - 6, the pdf's of α, θ and ϵ are uniquely determined by the joint distribution $(Y_{11}, Y_{12}, Y_{21}, Y_{22})$.*

Proof. We use Lemma 1 three times. First, denote $Z_1 = \alpha + \theta_1$ and $Z_2 = \alpha + \theta_2$. Lemma 1 implies that the joint distribution (Y_{11}, Y_{21}) uniquely determines the pdf of Z_1 and ϵ and

the joint distribution (Y_{12}, Y_{22}) uniquely determines the pdf of Z_2 and ϵ. Now letting the characteristic function of (Y_{11}, Y_{12}) be denoted as $\psi_{Y_{11}Y_{12}}(t_1, t_2)$, we have

$$\begin{aligned}\psi_{Y_{11}Y_{12}}(t_1, t_2) &= \mathbb{E}[\exp[\mathtt{i}(t_1(Z_1 + \epsilon_{11}) + t_2(Z_2 + \epsilon_{12}))]] \\ &= \phi_{Z_1Z_2}(t_1, t_2)\phi_\epsilon(t_1)\phi_\epsilon(t_2),\end{aligned}$$

where $\phi_{Z_1Z_2}(\cdot, \cdot)$ is the characteristic function of the random vector (Z_1, Z_2). The second equality holds under Assumption 4.

Since we have already identified the characteristic function ϕ_ϵ, the characteristic function of (Z_1, Z_2) is therefore identified. Now apply Lemma 1 again on

$$\begin{aligned}Z_1 &= \alpha + \theta_1 \\ Z_2 &= \alpha + \theta_2\end{aligned}$$

to identify the densities of α and θ. ■

Lemma 2 applies to the more general teacher value-added model with classroom shocks:

$$y_{ijt} = \alpha_j + \theta_{jt} + \epsilon_{ijt},$$

where i now indexes students, j indexes teachers and t indexes the academic year. With $E[\theta_{jt}] = 0$ and $E[\epsilon_{ijt}] = 0$ and assuming that α, θ_{jt}, and ϵ_{ijt} are mutually independent of each other, the pdf's of α, θ, and ϵ are nonparametrically identified.

B Construction of the Teacher Value-Added Sample

This appendix describes the construction of the final sample of students and teachers used for teacher VA estimation in both of our administrative data sets. Sample selection follows prior work (for instance, Chetty et al. (2014a,b)), the main requirements for inclusion in the sample being that the student has a valid score in a given subject both in the current and prior period, and can be matched to a teacher in that subject.

B.1 North Carolina

For North Carolina, we follow Clotfelter et al. (2006) and subsequent research using North Carolina data to construct our sample. We start with the entire enrollment history of students in North Carolina for grades 4-5 over the 1996-97 through 2010-11 school years and grade 3 over the 1996-97 through 2009-10 school years.[36] These data cover roughly 1.85 million students with 4.5 million student-year observations.

For demographics, we have information about parental education (six education groups, 1996-97 through 2005-06 only), economically disadvantaged status (1998-99 through 2010-11 only), ethnicity (six ethnic groups), gender, limited English status, disability status, academically gifted status and grade repetition. Besides the missing data in some years for parental education and economically disadvantaged status our demographic data cover over 99 percent of all student-year observations. Whenever demographic information is missing, we create a missing indicator for that variable.

We then make several sample restrictions. First, we drop the 1.37 million student-year observations we identify as having an invalid teacher. This is by far our biggest sample restriction and comes from the fact that we assign teachers to students based on the person recorded as proctoring the student's exam. To ensure the teacher proctoring the exam is the same as the classroom teacher, we confirm that the proctor is teaching a primary

[36] Our analysis is restricted to students in third through fifth grade since our data records the test proctor and the teacher recorded as the test proctor is typically the teacher who taught the students throughout the year in these grades. Data for grade 3 stops after 2008-09 because the grade 3 pretest was discontinued after that year. Grade 3 students in 2005-06 are also omitted due to a lack of the pre-test in the administrative data for that year.

grade mathematics and English class. If the teacher is not, we drop the observations as we are no longer confident of matching classes to teachers correctly. Second, we drop charter school classrooms and special education classrooms, leading to a loss of an additional 70,000 student-year observations. Third, we drop 16,000 observations where we lack data on teacher experience. Fourth, we exclude 380,000 observations that lack a valid current or lagged test score in that subject, with half of this loss coming from a lack of third grade mathematics pretest data in 2005-06 and third grade English pretest data in 2007-08 due to a statewide test update.[37] Fifth, we only include classes with more than seven but fewer than forty students with valid current and lagged test scores in that subject, creating a loss of 10,000 observations.[38] Our final sample consists of roughly 2.7 million student-year observations, covering 1.4 million students and 35,000 teachers.

B.2 Los Angeles Unified School District (LAUSD)

For the LAUSD, we start with the entire enrollment history of students in the district for grades 4-5 over the 2003-04 through 2012-13 and 2015-16 through 2016-17 school years and third grade from 2003-04 through 2012-13. These data cover roughly 800,000 students with 1.7 million student-year observations.

For demographics, we have information about parental education (five education groups), economically disadvantaged status, ethnicity (seven ethnic groups), gender, limited English status, age, and an indicator for skipping or repeating a grade. Demographic coverage is approximately one hundred percent for all demographic variables with the exception of parental education, which is missing for twenty-nine percent of the sample. Whenever parental education is missing, we create a missing indicator for that variable.

We then make several sample restrictions. First, we drop 100,000 student-year observations that cannot be matched to a teacher. Second, we drop 180,000 observations where we lack data on teacher experience. The data we drop here are over-represented in early years

[37] The third grade pretest is a test given to students at the start of third grade.

[38] As the last two restrictions are subject-specific, our sample for English value-added has 50,000 fewer student-year observations.

since we only have teacher experience data from 2007-08 onwards.[39] Third, we only include classes with more than seven but fewer than forty students with valid current and lagged test scores in that subject, losing 11,000 observations. Fourth, we exclude 70,000 observations that lack a valid current or lagged test score in that subject.[40] Our final sample is roughly 1.3 million student-year observations, covering roughly 660,000 million students and 11,000 teachers.

Constructing Value-Added: With both samples in hand, we construct VA estimates for each teacher by running the following regression:

$$y_{igt} = f_{1g}(y_{i,t-1}) + f_2(e_{j(i,g,t)}) + \phi_1 X_{igt} + \phi_2 \bar{X}_{c(i,g,t)} + v_j + \epsilon_{igt} \ .$$

We follow Chetty et al. (2014a,b) and parametrize the control function for lagged test scores $f_{1g}(y_{i,t-1})$ with a cubic polynomial in prior-year scores in mathematics and English and interact these cubics with the student's grade level. When prior test scores in the other subject are missing, we set the other subject prior score to zero and include an indicator for missing data in the other subject interacted with the controls for prior own-subject test scores.

We parametrize the control function for teacher experience $f_2(e_{j(i,g,t)})$ using dummies for years of experience from 0 to 5, with the omitted group being teachers with 6 or more years of experience. The student-level control vector X_{igt} consists of the respective demographic variables in each dataset. The class-level control vector $\bar{X}_{c(i,g,t)}$ includes (i) class size, (ii) cubics in class and school-grade means of prior-year test scores in mathematics and English each interacted with grade, (iii) class and school-year means of all the individual covariates, X_{igt}, and (iv) grade and year dummies.

[39] We assume teacher experience for teachers before 2007-08 is given by their experience in 2007-08 minus the number of years until 2007-08, but we cannot get teacher experience data for any teacher who left before 2007-08. We lose approximately 30% of observations in 2003-04, 25% in 2004-05, 17% in 2005-06, 10% in 2006-07. Every year thereafter we continue to lose about 3-5% of observations due missing values for teacher experience.

[40] As the last two restrictions are subject-specific, our sample for English VA has 4,000 fewer student-year observations.

C Bootstrapping the Standard Errors

This appendix discusses the bootstrapped standard errors generated for the policy evaluations. The uncertainty of the policy analysis, test score gains under teacher release or retaining policy, stems from the fact that the distribution of the teacher quality – either nonparametrically identified from the data or under the parametric assumption of Gaussian – requires estimation using the data. We apply the bootstrap method in Laird and Louis (1987) to construct standard errors for these policy evaluation estimates.

For policy estimates under general distribution F, the following steps describe the bootstrap procedure: (1) Draw a new independent sample of teacher quality of the same size as the original sample from distribution $\hat{F}$ and generate a bootstrap sample of the fixed effect estimates $y_j^{(b)}$ for $j = 1, 2, \ldots, n$ based on model (2.4); (2) Estimate the nonparametric MLE of $\hat{F}^{(b)}$ based on the bootstrap sample $\boldsymbol{y^{(b)}}$ and calculate the marginal and total test score gains based on $\hat{F}^{(b)}$. Repeat these steps for $B = 800$ times and calculate the standard error based on these bootstrap estimates of policy outcomes.

If the teacher quality is assumed to be unobserved and thus the cutoff for the bottom or top q percentile of quality needs to be found from the empirical quantiles of the estimates of the teacher value added, for each bootstrap distribution $\hat{F}^{(b)}$ conduct step (3): take an independent sample of teacher quality of size 40000 from distribution $\hat{F}^{(b)}$ and generate data based on model (2.4) with total class size equal to sixty. Construct the NPEB estimator of the value added and apply the policy of releasing or retaining based on empirical quantiles of the NPEB estimates of the teacher VA.

If we assume the quality distribution is normal, the following steps are taken to construct the bootstrap standard errors: (1) Draw a new independent sample of teacher quality of the same size as the original sample from $\mathcal{N}(0, \hat{\sigma}_\alpha^2)$ where $\hat{\sigma}_\alpha^2$ is the maximum likelihood estimator of variance of the normal distribution based on the respective dataset from North Carolina and the LAUSD. Then generate a bootstrap sample of the fixed effect estimates $y_j^{(b)}$ based on model (2.4); (2) Estimate $\hat{\sigma}_\alpha^{2(b)}$ using the maximum likelihood estimator applied to the bootstrapped sample $y_j^{(b)}$ and calculate the marginal and total test score gains based

on $\mathcal{N}(0, \hat{\sigma}_{\alpha}^{2(b)})$. If teacher quality is assumed to be unobserved, conduct step (3): take an independent sample of teacher quality of size 40000 from $\mathcal{N}(0, \hat{\sigma}_{\alpha}^{2(b)})$ and generate data based on model (2.4) with total class size equals to sixty. Construct the EB estimator of the value added and apply the policy of releasing or retaining teachers based on empirical quantiles of the EB estimates of teacher VA.

D Maximum Likelihood Estimation of Variance Parameters

D.1 Without Classroom Shocks

Our model without classroom shocks is specified as:

$$y_{ijt} = \alpha_j + \epsilon_{ijt}\,,$$

with i indexing students, j indexing teachers and t indexing the years for which teachers appear in the sample. We assume that $\epsilon_{ijt} \sim_{iid} \mathcal{N}(0, \sigma_\epsilon^2)$ and ϵ_{ijt} is independent of α_j. We have $i = 1, 2, \ldots, n_{jt}$, $j = 1, \ldots, J$ and $t = 1, \ldots, T_j$ (i.e., an unbalanced panel of teachers). Denote the teacher-year fixed effect $y_{jt} = \frac{1}{n_{jt}} \sum_i y_{ijt}$. The estimator commonly used in the literature is the following method of moment estimator proposed by Kane and Staiger (2008) under the additional assumption that $\alpha_j \sim N(0, \sigma_\alpha^2)$. In particular, they propose the following method of moment estimator for the variance parameters:

$$\hat{\sigma}_\alpha^2 = \widehat{\text{cov}}(y_{jt}, y_{jt-1})$$
$$\hat{\sigma}_\epsilon^2 = \widehat{V}(y_{ijt}) - \hat{\sigma}_\alpha^2$$

This is also the estimator used by Chetty et al. (2014a) for teacher VA without drift. The main shortcoming of the method of moment estimator for the variance parameters is that it requires all individual teachers to have shown up in the sample for at least 2 years; otherwise, they will be dropped from the covariance calculation. For North Carolina data, teachers who only appear for one year consist of around 30% of the whole sample. This induces a sample selection issue for the estimation of σ_α^2. We therefore propose the following maximum likelihood estimators for the variance parameters.

Maintaining a general distribution F for α, denote the vector $\vec{\boldsymbol{y}}_{\boldsymbol{jt}} = (y_{1jt}, y_{2jt}, \ldots, y_{n_{jt}jt})'$, the likelihood of observing residual test outcome $\vec{y}_{jt}$ for teacher j in period t can be written

as

$$
\begin{aligned}
L(\vec{\boldsymbol{y}}_{\boldsymbol{jt}}) &= \int \Big(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\Big)^{n_{jt}} \exp\Big(-\sum_i (y_{ijt} - y_{jt} + y_{jt} - \alpha_j)^2/2\sigma_\epsilon^2\Big) dF(\alpha_j) \\
&= \Big(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\Big)^{n_{jt}} \int \exp\Big(-\sum_i (y_{ijt} - y_{jt})^2/2\sigma_\epsilon^2\Big) \exp\Big(-\frac{(y_{jt} - \alpha_j)^2}{2\sigma_\epsilon^2/n_{jt}}\Big) dF(\alpha_j) \\
&= \frac{1}{\sqrt{n_{jt}}}\Big(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\Big)^{n_{jt}-1} \exp(-\sum_i (y_{ijt} - y_{jt})^2/2\sigma_\epsilon^2\Big) \int \frac{1}{\sqrt{2\pi\sigma_\epsilon^2/n_{jt}}} \exp\Big(-\frac{(y_{jt} - \alpha_j)^2}{2\sigma_\epsilon^2/n_{jt}}\Big) dF(\alpha_j) \\
&\equiv L_1(\vec{\boldsymbol{y}}_{\boldsymbol{jt}} | y_{jt}, \sigma_\epsilon^2) \int L_2(y_{jt}|\sigma_\epsilon^2, \alpha_j) dF(\alpha_j)
\end{aligned}
$$

When F is assumed to be the normal distribution with variance σ_α^2, then the second component involving the integral becomes

$$
\int L_2(y_{jt}|\sigma_\epsilon^2, \alpha_j) dF(\alpha_j) = \frac{1}{\sqrt{2\pi(\sigma_\alpha^2 + \sigma_\epsilon^2/n_{jt})}} \exp\Big(-\frac{y_{jt}^2}{2(\sigma_\alpha^2 + \sigma_\epsilon^2/n_{jt})}\Big) := \tilde{L}_2(y_{jt}|\sigma_\epsilon^2, \sigma_\alpha^2)
$$

Therefore, under the normality assumption, the maximum likelihood estimator for $(\sigma_\epsilon^2, \sigma_\alpha^2)$ can be obtained by maximizing $\prod_j \prod_t L_1(\vec{\boldsymbol{y}}_{\boldsymbol{jt}} | y_{jt}, \sigma_\epsilon^2) \tilde{L}_2(y_{jt}|\sigma_\epsilon^2, \sigma_\alpha^2)$ numerically. Unlike the method of moment estimator, here all individuals, including those only have one period of data, are accounted for.

When F is a general distribution not indexing by any parameters, we can obtain an estimator for σ_ϵ^2 through

$$
\hat{\sigma}_\epsilon^2 = \frac{\sum_j \sum_t \sum_i (y_{ijt} - y_{jt})^2}{\sum_j \sum_t (n_{jt} - 1)}.
$$

D.2 With Classroom Shocks

Our model with classroom shocks is specified as:

$$
y_{ijt} = \alpha_j + \theta_{jt} + \epsilon_{ijt}\,,
$$

with i indexing students, j indexing teachers and t indexing years for which teachers appear in the sample. We assume that $\theta_{jt} \sim \mathcal{N}(0, \sigma_\theta^2)$ and $\epsilon_{ijt} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and mutual independence between α_j, θ_{jt} and ϵ_{ijt}. We again have $i = 1, 2, \ldots, n_{jt}$, $j = 1, \ldots, J$ and $t = 1, \ldots, T_j$ (i.e.,

an unbalanced panel of teachers).

The corresponding method of moment estimator proposed by Kane and Staiger (2008) is

$$\begin{aligned}\hat{\sigma}_\epsilon^2 &= \hat{V}(y_{ijt} - y_{jt}) \\ \hat{\sigma}_\alpha^2 &= \widehat{\text{cov}}(y_{jt}, y_{jt-1}) \\ \hat{\sigma}_\theta^2 &= \hat{V}(y_{ijt}) - \hat{\sigma}_\epsilon^2 - \hat{\sigma}_\alpha^2\end{aligned}$$

Again, the method of moment estimator excludes individual teachers who appear for only one period in the sample. As an alternative, we propose the following maximum likelihood estimator for the variance parameters.

Parametric EB: Under PEB, the VA of teacher j is assumed to be distributed according to $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$. Denoting the vector $\vec{\boldsymbol{y}}_{\boldsymbol{jt}} = (y_{1jt}, \ldots, y_{n_{jt}jt})'$, the likelihood of $\vec{\boldsymbol{y}}_{\boldsymbol{jt}}$ can be written as

$$L(\vec{\boldsymbol{y}}_{\boldsymbol{jt}}) = \int (2\pi)^{-n_{jt}/2} |\text{det}\Sigma|^{-1/2} exp(-\frac{1}{2}(\vec{\boldsymbol{y}}_{\boldsymbol{jt}} - \alpha_j)^\top \Sigma^{-1}(\vec{\boldsymbol{y}}_{\boldsymbol{jt}} - \alpha_j)) dF(\alpha_j)\,,$$

where $\Sigma = \sigma_\epsilon^2 I + \sigma_\theta^2 \mathbf{1}_{n_{jt}} \mathbf{1}'_{n_{jt}}$ with I being an identity matrix of dimension $n_{jt} \times n_{jt}$ and $\mathbf{1}_n$ is a vector of 1's with length n. Some algebra shows that

$$\text{det}\Sigma = \left[n_{jt}\sigma_\theta^2 + \sigma_\epsilon^2\right](\sigma_\epsilon^2)^{n_{jt}-1}\,,$$

and

$$\Sigma^{-1} = \frac{1}{\sigma_\epsilon^2} I - \frac{\sigma_\theta^2}{(\sigma_\epsilon^2 + n_{jt}\sigma_\theta^2)\sigma_\epsilon^2} \mathbf{1}_{n_{jt}} \mathbf{1}'_{n_{jt}}\,.$$

Now, defining $y_{jt} := \frac{1}{n_{jt}} \sum_i y_{ijt}$ gives us

$$
\begin{aligned}
&(\vec{\boldsymbol{y}}_{\boldsymbol{jt}} - \alpha_j)' \Sigma^{-1} (\vec{\boldsymbol{y}}_{\boldsymbol{jt}} - \alpha_j) \\
&= \frac{1}{\sigma_\epsilon^2} \sum_i (y_{ijt} - \alpha_j)^2 - \frac{\sigma_\theta^2}{(\sigma_\epsilon^2 + \sigma_\theta^2 n_{jt}) \sigma_\epsilon^2} \Big(\sum_i (y_{ijt} - \alpha_j) \Big)^2 \\
&= \frac{1}{\sigma_\epsilon^2} \sum_i (y_{ijt} - y_{jt} + y_{jt} - \alpha_j)^2 - \frac{\sigma_\theta^2}{(\frac{\sigma_\epsilon^2}{n_{jt}} + \sigma_\theta^2) \frac{\sigma_\epsilon^2}{n_{jt}}} \Big(y_{jt} - \alpha_j \Big)^2 \\
&= \frac{1}{\sigma_\epsilon^2} \sum_i (y_{ijt} - y_{jt})^2 + \frac{1}{\frac{\sigma_\epsilon^2}{n_{jt}} + \sigma_\theta^2} (y_{jt} - \alpha_j)^2 ,
\end{aligned}
$$

and then the likelihood of observing the vector $\vec{\boldsymbol{y}}_{\boldsymbol{jt}}$ for teacher j at period t (conditional on α_j) becomes

$$
\begin{aligned}
L(\vec{\boldsymbol{y}}_{\boldsymbol{jt}} | \alpha_j) &= (2\pi)^{-n_{jt}/2} |\det \Sigma|^{-1/2} \exp\Big(- \frac{\sum_i (y_{ijt} - y_{jt})^2}{2\sigma_\epsilon^2} \Big) \exp\Big(- \frac{(y_{jt} - \alpha_j)^2}{2(\frac{\sigma_\epsilon^2}{n_{jt}} + \sigma_\theta^2)} \Big) \\
&= (2\pi)^{-\frac{n_{jt}-1}{2}} |\det \Sigma|^{-1/2} \exp\Big(- \frac{\sum_i (y_{ijt} - y_{jt})^2}{2\sigma_\epsilon^2} \Big) \Big(\frac{\sigma_\epsilon^2}{n_{jt}} + \sigma_\theta^2 \Big)^{1/2} \frac{1}{\sqrt{2\pi(\frac{\sigma_\epsilon^2}{n_{jt}} + \sigma_\theta^2)}} \exp\Big(- \frac{(y_{jt} - \alpha_j)^2}{2(\frac{\sigma_\epsilon^2}{n_{jt}} + \sigma_\theta^2)} \Big) \\
&= (2\pi)^{-\frac{n_{jt}-1}{2}} n_{jt}^{-\frac{1}{2}} (\sigma_\epsilon^2)^{-\frac{n_{jt}-1}{2}} \exp\Big(- \frac{\sum_i (y_{ijt} - y_{jt})^2}{2\sigma_\epsilon^2} \Big) \frac{1}{\sqrt{2\pi(\frac{\sigma_\epsilon^2}{n_{jt}} + \sigma_\theta^2)}} \exp\Big(- \frac{(y_{jt} - \alpha_j)^2}{2(\frac{\sigma_\epsilon^2}{n_{jt}} + \sigma_\theta^2)} \Big) .
\end{aligned}
$$

Note that $y_{jt} | \alpha_j \sim N(\alpha_j, \sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}})$ and so the second piece of the likelihood is of itself a proper likelihood for y_{jt} conditional on α_j and the first piece of the likelihood does not depend on α_j or σ_θ^2. If $\alpha_j \sim N(0, \sigma_\alpha^2)$, then $y_{jt} \sim N(0, \sigma_\alpha^2 + \sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}})$. Then the marginal likelihood of all the data (unconditional on α_j) becomes

$$
L = \prod_j \prod_t \Big\{ (2\pi)^{-\frac{n_{jt}-1}{2}} n_{jt}^{-\frac{1}{2}} (\sigma_\epsilon^2)^{-\frac{n_{jt}-1}{2}} \exp\Big(- \frac{\sum_i (y_{ijt} - y_{jt})^2}{2\sigma_\epsilon^2} \Big) \frac{1}{\sqrt{2\pi(\sigma_\theta^2 + \sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n_{jt}})}} \exp\Big(- \frac{y_{jt}^2}{2(\sigma_\theta^2 + \sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n_{jt}})} \Big) \Big\}
$$

The maximum likelihood estimator for $(\sigma_\alpha^2, \sigma_\theta^2, \sigma_\epsilon^2)$ can be solved by maximizing L numerically.

NPEB: Under NPEB, we have that $\alpha_j \sim F$. We start by estimating σ_ϵ^2 from the first piece of the likelihood over (j, t), that is

$$
\hat{\sigma}_\epsilon^2 = \underset{\sigma_\epsilon^2}{\operatorname{argmax}} \sum_j \sum_t - \frac{n_{jt} - 1}{2} \log \sigma_\epsilon^2 - \frac{\sum_i (y_{ijt} - y_{jt})^2}{2\sigma_\epsilon^2} ,
$$

which leads to

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_j \sum_t \sum_i (y_{ijt} - y_{jt})^2}{\sum_j \sum_t (n_{jt} - 1)} \ .$$

Now to estimate σ_θ^2, consider the model $y_{jt}|\alpha_j \sim N(\alpha_j, \sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}})$. The likelihood for the vector $(y_{j1}, \ldots, y_{jT_j})$ can be written as

$$L(y_{j1}, \ldots, y_{jT_j}|\alpha_j) = \Big[\prod_{t=1}^{T_j} \frac{1}{\sqrt{2\pi(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}})}}\Big] \exp\Big(-\frac{1}{2}\sum_t \frac{(y_{jt} - \alpha_j)^2}{\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}}}\Big) \ .$$

Letting $\nu_{jt} = \sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}}$, define

$$y_j = \frac{\sum_t \frac{y_{jt}}{\nu_{jt}}}{\sum_t \frac{1}{\nu_{jt}}} \ .$$

We then have $y_j|\alpha_j \sim N(\alpha_j, \frac{1}{\sum_t \frac{1}{\nu_{jt}}})$ and the likelihood of $L(y_{j1}, \ldots, y_{jT_j}|\alpha_j)$ factorizes into

$$\Big[\prod_{t=1}^{T_j} \frac{1}{\sqrt{2\pi(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}})}}\Big] \exp\Big(-\frac{1}{2}\sum_t \frac{(y_{jt} - y_j)^2}{\nu_{jt}}\Big) \sqrt{2\pi \frac{1}{\sum_t \frac{1}{\nu_{jt}}}} \frac{1}{\sqrt{2\pi \frac{1}{\sum_t \frac{1}{\nu_{jt}}}}} \exp\Big(-\frac{1}{2}(y_j - \alpha_j)^2 \sum_t \frac{1}{\nu_{jt}}\Big) \ ,$$

where the second piece forms the density of y_j conditional on α_j. We estimate σ_θ^2 by maximizing the following likelihood

$$\prod_j \Big\{\Big[\prod_{t=1}^{T_j} \frac{1}{\sqrt{2\pi(\sigma_\theta^2 + \frac{\hat{\sigma}_\epsilon^2}{n_{jt}})}}\Big] \exp\Big(-\frac{1}{2}\sum_t \frac{(y_{jt} - y_j)^2}{\sigma_\theta^2 + \frac{\hat{\sigma}_\epsilon^2}{n_{jt}}}\Big) \sqrt{2\pi \frac{1}{\sum_t \frac{1}{\sigma_\theta^2 + \frac{\hat{\sigma}_\epsilon^2}{n_{jt}}}}}\Big\} \ .$$

There is no closed-form solution for $\hat{\sigma}_\theta^2$, but numerical estimates can be easily obtained.

E Specification Test for Normality

We propose the following specification test for normality. Suppose the data are generated from the model

$$y_j = \alpha_j + \epsilon_j, \quad j = 1, 2, \ldots, n\,,$$

with $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ and where σ_j^2 is known. We are interested in testing the hypothesis that $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$. One natural diagnostic test is the likelihood ratio test, with the test statistic given by

$$L_n = 2\Big(\sup_{F \in \mathcal{F}} \ell_n(F) - \sup_{\sigma_\alpha^2} \ell_n(\sigma_\alpha^2)\Big)\,,$$

where $\mathcal{F}$ is the set of probability measures on the domain of α, $\ell_n(F)$ is the likelihood of the sample $\{\bar{v}_1, \ldots, \bar{v}_n\}$ with $\alpha \sim F$, and $\ell_n(\sigma_\alpha^2)$ is the likelihood of the sample with $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$. To obtain a critical value for the test based on L_n, we use the parametric bootstrap, drawing on McLachlan (1987) and Gu et al. (2018). This involves the following steps:

1. Compute $\hat{\sigma}_\alpha^2$ as the maximizer of $\ell_n(\sigma_\alpha^2)$.

2. For $b = 1, \ldots, B$, generate data $\alpha_1^{(b)}, \ldots, \alpha_n^{(b)}$ from $\mathcal{N}(0, \hat{\sigma}_\alpha^2)$.

3. For $b = 1, \ldots, B$, generate data $y_j^{(b)}$ from $\mathcal{N}(\alpha_j^{(b)}, \sigma_j^2)$ for $j = 1, 2, \ldots, n$.

4. For $b = 1, \ldots, B$, denote by $L_{n,b}$ the test statistic L_n computed from the sample $y_1^{(b)}, \ldots, y_n^{(b)}$. Compute the τ-quantile $q_{n,\tau}$ of $L_{n,1}, \ldots, L_{n,B}$.

The likelihood ratio test statistic computed from the data takes the form $L_n = 2(\ell_n(\hat{F}) - \ell_n(\hat{\sigma}_\alpha^2)$ where $\hat{F}$ is the NPMLE defined in the main text and σ_α^2 is the maximum likelihood estimator under the assumption that $\alpha_j \sim N(0, \sigma_\alpha^2)$. Details are given in Appendix D in the paper. We reject the null hypothesis of a normal distribution for the teacher quality α at level τ when L_n exceeds the bootstrap-based critical value $q_{n,1-\tau}$.

We report the size and power performance of the proposed parametric bootstrap test in Table E.1 below with the following data generating process: Fix the sample size at $n = 1000$,

and for a grid values of $h \in \{0, 0.4, 0.6, 0.8, 1\}$, sample individual α_j's from the following three-component normal distribution:

$$0.025\mathcal{N}(-h, \theta_h) + 0.95\mathcal{N}(0, \theta_h) + 0.025\mathcal{N}(h, \theta_h)$$

with $\theta_h = 0.1 - 0.05h^2$. The design of θ_h is such that the variance of α is always 0.1; this is roughly the variance of the teacher effects in the LAUSD data. When $h = 0$, the latent effect α_j follows a normal distribution, and the bootstrap test should reject with probability equal to nominal size. As the magnitude of h increases, we deviate from the normal distribution, and the parametric bootstrap test should be able to detect this deviation from the null hypothesis of normality and reject with a higher probability. Conditional on α_j, the data y_j is generated from a normal distribution with mean α_j and variance σ_j^2, where the σ_j's are generated from a random sample of size 1000 from the inverse gamma distribution with parameters $(6, 0.05)$, in order to capture teacher heterogeneity. These parameters are chosen so that the distribution of σ_j^2 mimics those for the individual variances in the LAUSD data.

Results in Table E.1 are based on bootstrap sample size $B = 500$ and 500 simulation repetitions. Table E.1 shows that the parametric bootstrap test controls size well for $h = 0$, and that the power increases quickly as h increases.

Table E.1: Size and Power Performance of the Parametric Bootstrap Test for Normality

	$\tau = 10\%$	$\tau = 5\%$	$\tau = 1\%$
$h = 0$	0.116	0.058	0.01
$h = 0.4$	0.148	0.082	0.028
$h = 0.6$	0.57	0.442	0.234
$h = 0.8$	1	1	0.99
$h = 1$	1	1	1

Notes: τ measures the nominal sizes fixed at $10, 5, 1\%$ and we report the proportion of rejection out of 500 simulation repetitions for different values of h and τ.

We apply the parametric bootstrap likelihood ratio test of normality for both NC and LAUSD data. For the NC data, the likelihood ratio test statistic L_n is 1326.3 with the

corresponding bootstrap critical values at $(1-\tau) \in \{90\%, 95\%, 99\%\}$ being respectively $\{58.45, 61.67, 70.4\}$, which implies that the normality hypothesis is significantly rejected at 1% level. For LAUSD data, the likelihood ratio test statistics L_n is 667.5 and the corresponding bootstrap critical values at $\tau \in \{90\%, 95\%, 99\%\}$ being respectively $\{73.8, 78.1, 90.1\}$ and hence we also reject the null hypothesis of normality at 1% level.

Other tests for normality is also possible. For instance, if α indeed follows a normal distribution $N(0, \sigma_\alpha^2)$, then the logrithm of its characteristic function takes the form

$$\log \phi_\alpha(t) = -t^2/\sigma_\alpha^2 \ ,$$

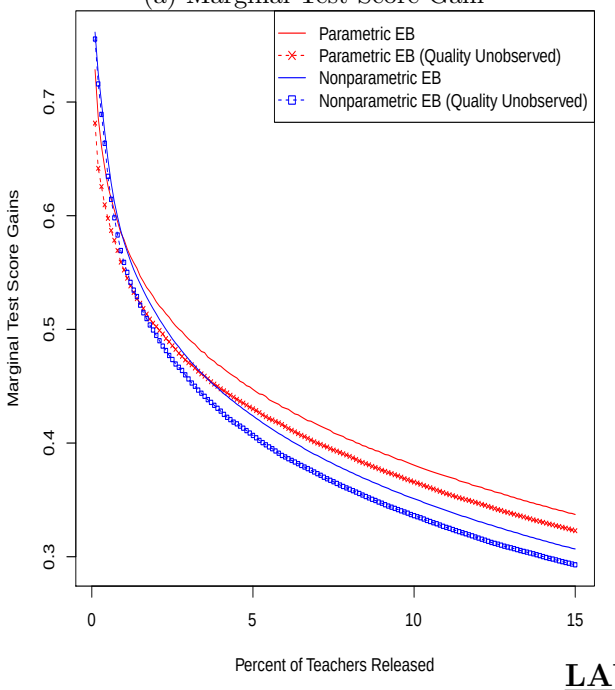
which implies that the first-order derivative with respect to t is of the form $-t/\sigma_\alpha^2$ which is a linear function of t. Since the distribution of α is identified (as established in Theorem 1), we can construct a consistent estimator for $\phi_\alpha(t)$ and inspect linearity of the derivative of its logarithm transformation. Another specification test is also proposed in Bonhomme and Weidner (2019). We leave to future research a power comparison involving different specification tests.

F Appendix Figures and Tables

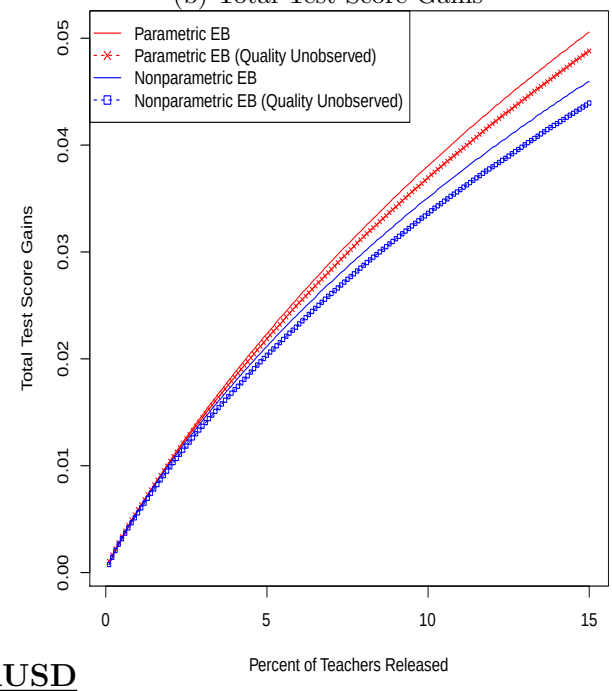
Figure F.1: Gains from Replacing Bottom q Percentile of Teachers when VA is Estimated

North Carolina

(a) Marginal Test Score Gain

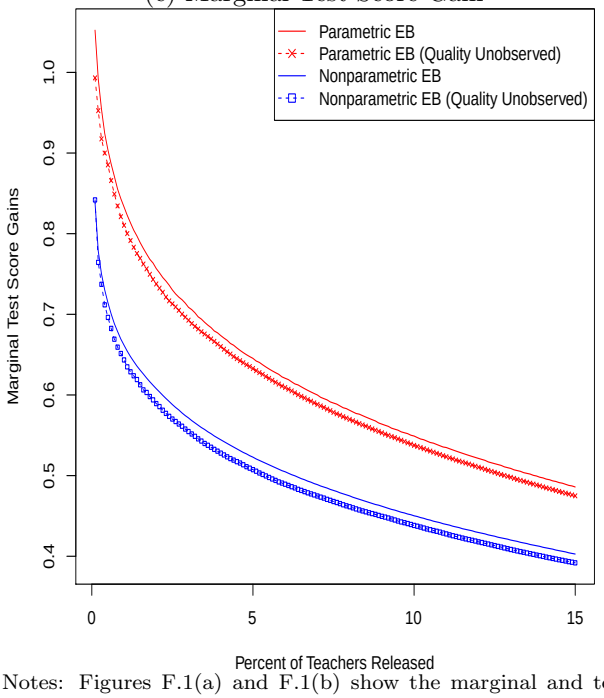


(b) Total Test Score Gains

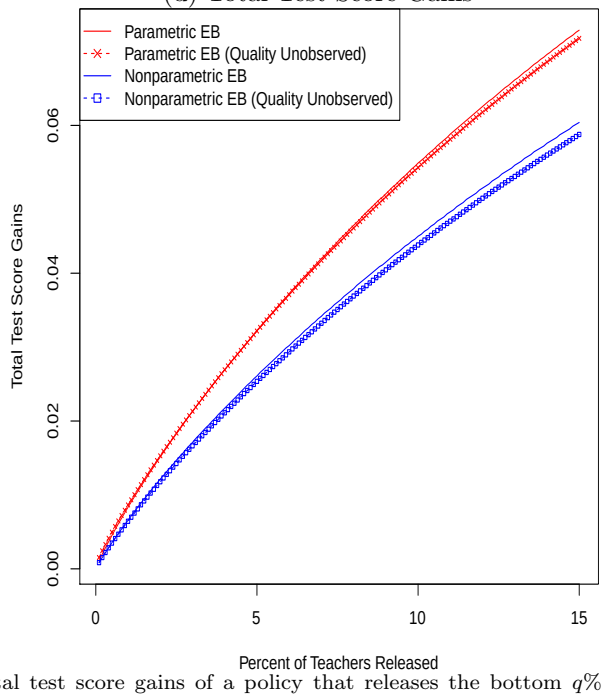


LAUSD

(c) Marginal Test Score Gain



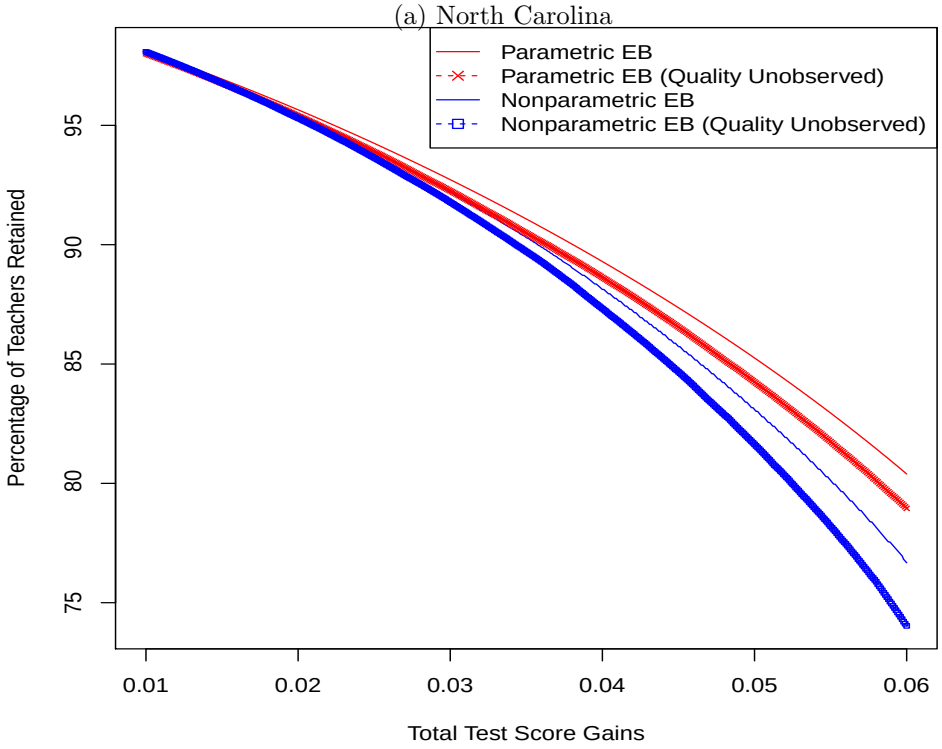
(d) Total Test Score Gains



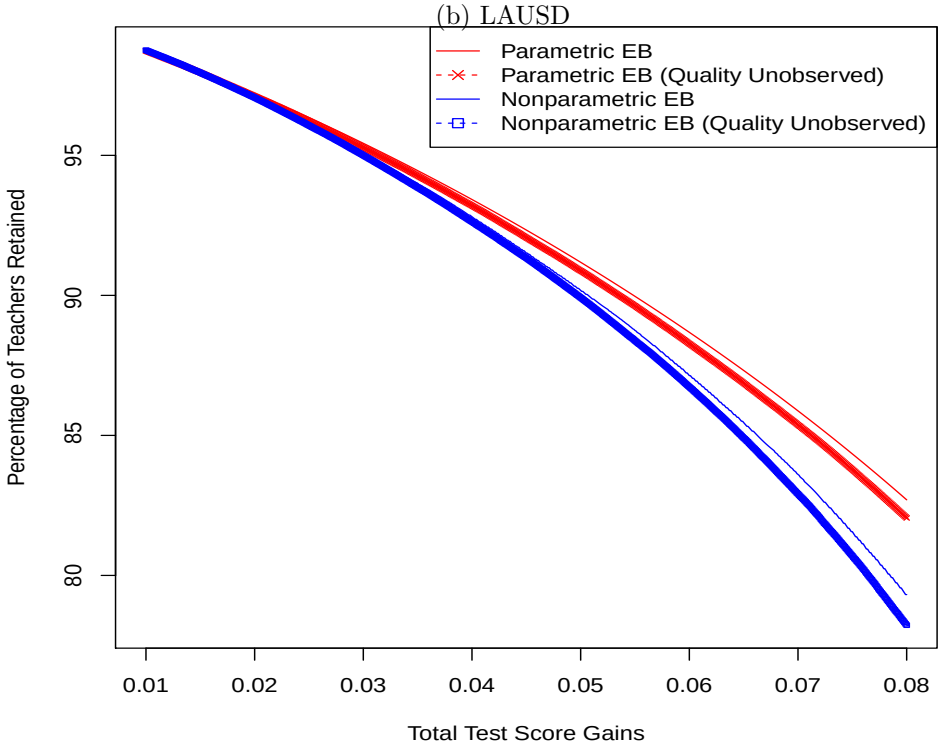
Notes: Figures F.1(a) and F.1(b) show the marginal and total test score gains of a policy that releases the bottom q% of teachers in North Carolina, while Figures F.1(c) and F.1(d) do the same for the LAUSD. The solid lines indicate the policy gains expected under the PEB and NPEB methodology when true teacher VA is observed and are identical to those presented in Figure 5. The dashed lines represent the policy gains when VA is estimated. These gains are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. Details of the simulation are provided in Section 7.1. Results in the figures are the same as those reported in Tables 4(a) and 4(b).

Figure F.2: Test Scores Gains from Retaining Top $1-q$ Percentile Teachers when VA is Estimated

(a) North Carolina



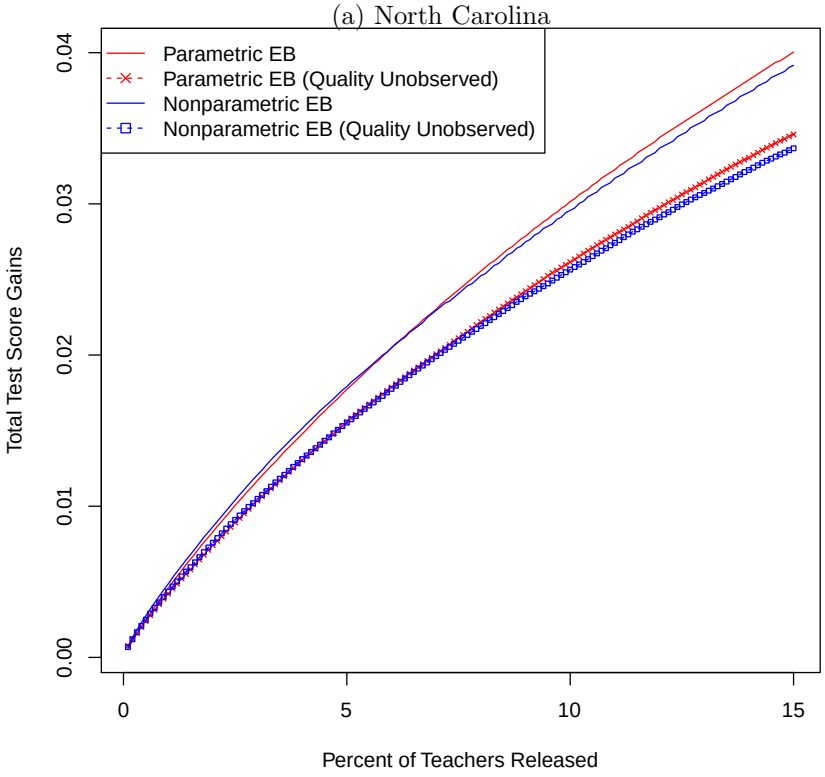
(b) LAUSD



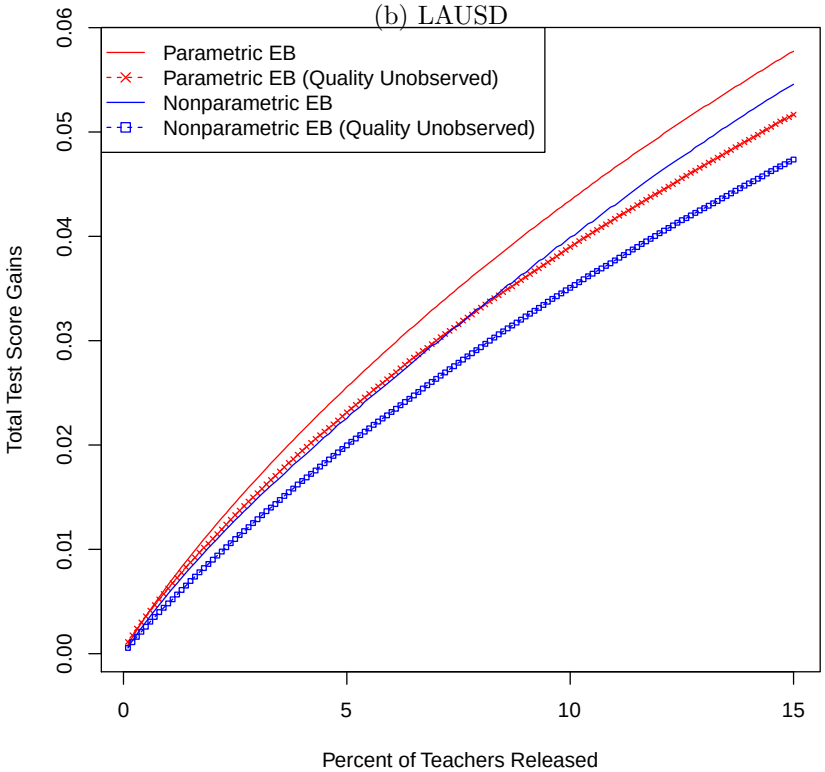
Notes: Figures F.2(a) and F.2(b) display the total test score gains from retaining teachers above the $1-q^{th}$ percentile of the value-added distribution in North Carolina and the LAUSD, respectively. The solid lines indicate the policy gains expected under the PEB and NPEB methodology when true teacher VA is observed. The dashed lines represent the policy gains when VA is estimated. These gains are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. Details of the simulation are provided in Section 7.2. Results in the figures are identical to those reported in Tables 5(a) and 5(b).

Figure F.3: Classroom Shocks Model: Test Scores Gains from Replacing Bottom q Percentile of Teachers when VA is Estimated

(a) North Carolina



(b) LAUSD



Notes: Figures F.3(a) and F.3(b) show the total test score gains under the classroom shocks model presented in equation (7.8) of a policy that releases the bottom q% of teachers in North Carolina and LAUSD, respectively. The solid lines indicate the policy gains expected under the PEB and NPEB methodology when true teacher value-added is observed. The dashed lines represent the policy gains when value-added is estimated. These gains are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. Details of the simulation are provided in Section 7.1. Results in the figures are identical to those reported in Tables F.2(a) and F.2(b).

Table F.1(a): Simulation (True Distribution is Normal)

	Homogeneous Class Sizes (Total Class Size of 20)				Heterogeneous Class Sizes (Total Class Size 20-40)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Mean Squared Error	10.81	10.87	10.81	12.50	11.03	11.07	11.03	12.84
Teacher ranked in bottom (top) 5% when true VA above (below) 5% (Type I error)								
(Type I error=Type II error by definition)								
Bottom 5%	152.5	152.5	152.5	152.5	153.5	153.6	153.5	153.9
Top 5%	152.6	152.6	152.6	152.6	153.6	153.5	153.6	154.1

Table F.1(b): Simulation (True Distribution is Mixed Normal)

	Homogeneous Class Sizes (Total Class Size of 20)				Heterogeneous Class Sizes (Total Class Size 20-40)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Mean Squared Error	9.08	9.14	10.82	12.51	7.14	7.18	8.29	9.36
Teacher ranked in bottom (top) 5% when true VA above (below) 5% (Type I error)								
(Type I error=Type II error by definition)								
Bottom 5%	126.9	126.9	126.9	126.9	111.8	111.9	111.7	113.7
Top 5%	126.0	126.0	126.0	126.0	111.4	111.5	112.0	114.0

Table F.1(c): Simulation (True Distribution is Chi-Squared)

	Homogeneous Class Sizes (Total Class Size of 20)				Heterogeneous Class Sizes (Total Class Size 20-40)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Mean Squared Error	7.44	7.45	10.82	12.51	5.79	5.79	8.30	9.37
Teacher ranked in bottom (top) 5% when true VA above (below) 5% (Type I error)								
(Type I error=Type II error by definition)								
Bottom 5%	431.3	431.3	431.3	431.3	425.5	425.6	427.5	428.6
Top 5%	66.4	66.4	66.4	66.4	56.8	56.9	57.4	57.0

Notes: This table adds type I and type II error rates to Tables 1(a), 1(b) and 1(c); given this mean squared error is the same as in those tables. The table uses simulation to compare the performance of four estimators when the distribution of teacher value-added follows a normal, mixed normal and a chi-squared distribution, respectively. Specifically, Table 1(a) has teacher value-added being normally distributed with mean zero and variance 0.05 (i.e., $F \sim \mathcal{N}(0, 0.08)$), while Table 1(b) has true teacher quality following $F \sim 0.98\mathcal{N}(0, 0.03) + 0.01\mathcal{N}(-1, 0.03) + 0.01\mathcal{N}(1, 0.03)$ (as in the example given by equation (2.8)). The normal and mixed normal distributions have the same mean and variance to create a suitable comparison. Teacher value-added follows $F \sim \chi^2_1$ in Table 1(c). The infeasible estimator is the optimal estimator given that the true distribution is known to the econometrician (which is infeasible as it is unknown in practice), the nonparametric emiprical Bayes (NPEB) estimator which nonparametrically estimates the underlying distribution, the parametric empirical Bayes (PEB) estimator which assumes that the underlying distribution is normal, and the fixed effect (FE) estimator which applies no empirical Bayes shrinkage. The simulation averages results from 500 repetitions with 10,000 individual teachers setting $\sigma^2_\epsilon = 0.025$. Results are reported for homogeneous class sizes where every teacher has a class size of twenty students and heterogeneous class sizes where class sizes of teachers are a random draw from the set $\{20, 40\}$ with equal probability. Note that teacher rankings are identical for the three methods under homogeneous class sizes. Only Type I error (teacher ranked in bottom (top) 5% when true VA above (below) 5%) is reported as it is identical to that of Type II error (teacher ranked above (below) 5% when true VA below (above) 5%).

Table F.2(a): Classroom Shocks Model: Test Scores Gains of Policy Releasing Bottom q% Teachers (True Value-Added Observed)

	North Carolina Data		LAUSD Data	
% Teachers Released (q)	Test Gain under F (NPEB) (1)	Test Gain under Normal (PEB) (2)	Test Gain under F (NPEB) (3)	Test Gain under Normal (PEB) (4)
1	0.0049	0.0046	0.0058	0.0066
	(0.0001)	(0.0000)	(0.0002)	(0.0001)
3	0.0121	0.0117	0.0149	0.0168
	(0.0002)	(0.0001)	(0.0004)	(0.0001)
5	**0.0179**	**0.0177**	**0.0225**	**0.0256**
	(0.0003)	(0.0001)	(0.0005)	(0.0002)
7	0.0230	0.0230	0.0297	0.0332
	(0.0004)	(0.0001)	(0.0006)	(0.0003)
9	0.0275	0.0279	0.0365	0.0402
	(0.0004)	(0.0001)	(0.0006)	(0.0003)

Table F.2(b): Classroom Shocks Model: Test Scores Gains of Policy Releasing Bottom q% Teachers (True Value-Added Unobserved)

	North Carolina Data		LAUSD Data	
% Teachers Released (q)	Test Gain under F (NPEB) (1)	Test Gain under Normal (PEB) (2)	Test Gain under F (NPEB) (3)	Test Gain under Normal (PEB) (4)
1	0.0044	0.0042	0.0048	0.0063
	(0.0003)	(0.0001)	(0.0001)	(0.0001)
3	0.0105	0.0104	0.0129	0.0154
	(0.0004)	(0.0002)	(0.0003)	(0.0002)
5	**0.0155**	**0.0155**	**0.0200**	**0.0231**
	(0.0005)	(0.0003)	(0.0004)	(0.0003)
7	0.0199	0.0201	0.0263	0.0300
	(0.0006)	(0.0004)	(0.0004)	(0.0004)
9	0.0239	0.0242	0.0323	0.0361
	(0.0007)	(0.0004)	(0.0005)	(0.0004)

Notes: Table F.2(a) and displays the estimated gains using our model that includes classroom shocks (see equation (7.8)) in mathematics scores in terms of student level standard deviations of a policy that releases the bottom q% of teachers and replaces them with mean quality teachers when true teacher quality is observed by the policymaker. 'Test score gain under F' reports the test score gain of the policy when teacher quality is distributed according the distribution F – nonparametrically estimated using equation (3.1) – and applying the NPEB estimator to calculate value-added. 'Test score gain under normal' reports the test score gain when teacher quality is normally distributed and the PEB estimator is used to calculate teacher value-added. Table F.2(b) repeats the exercise when the true teacher quality is unobserved to the policymaker and so teacher releases are based on estimated (rather than true) value-added. These gains are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. The bolded line indicates the widely-analyzed release bottom five percent teachers policy. Reported policy gains for both tables are identical to those shown in Figure F.3. Standard errors are calculated using a bootstrap (as described in Appendix C).

Table F.3(a): Classroom Shocks Model – Test Scores Gains of Policy Retaining 'Top $1-q$ Percentile' Teachers (True VA Observed)

	North Carolina Data		LAUSD Data	
% Teachers Retained $(1-q)$	Test Gain under F (NPEB) (1)	Test Gain under Normal (PEB) (2)	Test Gain under F (NPEB) (3)	Test Gain under Normal (PEB) (4)
1	0.0056	0.0046	0.0079	0.0066
	(0.0001)	(0.0000)	(0.0002)	(0.0001)
3	0.0134	0.0117	0.0193	0.0168
	(0.0002)	(0.0001)	(0.0004)	(0.0001)
5	0.0199	0.0177	0.0286	0.0256
	(0.0003)	(0.0001)	(0.0005)	(0.0002)
7	0.0254	0.0230	0.0368	0.0332
	(0.0003)	(0.0001)	(0.0006)	(0.0003)
9	0.0303	0.0279	0.0442	0.0402
	(0.0004)	(0.0001)	(0.0007)	(0.0003)

Table F.3(b): Classroom Shocks Model – Test Scores Gains of Policy Retaining 'Top $1-q$ Percentile' Teachers (True VA Unobserved)

	North Carolina Data		LAUSD Data	
% Teachers Retained $(1-q)$	Test Gain under F (NPEB) (1)	Test Gain under Normal (PEB) (2)	Test Gain under F (NPEB) (3)	Test Gain under Normal (PEB) (4)
1	0.0053	0.0036	0.0074	0.0054
	(0.0003)	(0.0001)	(0.0003)	(0.0001)
3	0.0123	0.0097	0.0179	0.0145
	(0.0004)	(0.0002)	(0.0004)	(0.0002)
5	0.0178	0.0149	0.0265	0.0221
	(0.0005)	(0.0003)	(0.0005)	(0.0003)
7	0.0227	0.0195	0.0340	0.0291
	(0.0006)	(0.0004)	(0.0006)	(0.0004)
9	0.0268	0.0237	0.0407	0.0353
	(0.0007)	(0.0004)	(0.0007)	(0.0004)

Notes: Table F.3(a) displays the estimated test score gains (in student-level SDs) of a policy that retains the top q% of teachers, rather than having them leave teaching and be replace by a mean quality teacher, when true teacher quality is observed by the policymaker. It uses a variant of our model that includes classroom shocks (see equation (7.8)) in mathematics scores. The columns headed 'Test Score Gain under F' report the policy gain when teacher quality is distributed according the distribution F, estimated nonparametrically using equation (3.1) and applying the NPEB estimator to calculate value-added. The columns headed 'Test Score Gain under Normal' report the policy gain when teacher quality is normally distributed and the PEB estimator is used to calculate teacher value-added. Table F.3(b) repeats the exercise when the true teacher quality is unobserved to the policymaker and so teacher retention is based on estimated (rather than true) value-added. These gains are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming class sizes of twenty. Standard errors are calculated using a bootstrap (as described in Appendix C).