University of Toronto
Department of Economics

Working Paper 689

A Nonparametric Method for Estimating Teacher Value-Added

By Michael Gilraine, Jiaying Gu and Robert McMillan

February 13, 2021

A Nonparametric Method for Estimating Teacher Value-Added*

Michael Gilraine, New York University
Jiaying Gu, University of Toronto
Robert McMillan, University of Toronto and NBER

February 11, 2021

Abstract

This paper proposes a computationally feasible nonparametric methodology for estimating teacher value-added. Our estimator, drawing on Robbins (1956), permits the unobserved teacher value-added distribution to be estimated directly, rather than assuming normality as is standard. Simulations indicate the estimator performs very well regardless of the true distribution, even in moderately-sized samples. Implementing our method in practice using two large-scale administrative datasets, the estimated teacher value-added distributions depart from normality and differ from each other. Further, compared with widely-used parametric estimates, we show our nonparametric estimates can make a significant difference to teacher-related policy calculations, in both short and longer terms.

JEL Codes: C11, H75, I21, J24.

Keywords: Teacher Value-Added, Nonparametric Empirical Bayes, Education Policy, Teacher Release Policy

*This is a revised version of NBER Working Paper 27094. We would like to thank Roger Koenker for many helpful discussions, and Joe Altonji, Raj Chetty, Michael Dinerstein, John Friedman, Nikolaos Ignatiadis, Chris Taber, Sergio Urzua and seminar participants at the NBER Summer Institute, Yale University, the University of Maryland, Western, Hunter College, 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. All remaining errors are our own. Contact emails: Gilraine, mike.gilraine@nyu.edu; Gu, jiaying.gu@utoronto.ca; McMillan, 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, researchers have proposed influential fixed effects methods intended to capture a teacher's underlying quality, taking advantage of large-scale matched student-teacher datasets that are increasingly accessible – see pioneering studies by Rockoff (2004) and Rivkin, Hanushek, and Kain (2005). In turn, these methods have prompted the development of value-added ('VA') estimators for measuring teacher quality – Koedel, Mihaly, and Rockoff (2015) provide a recent review – with teacher VA estimates now featuring widely 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 estimates in high-stakes decision making raises important challenges. Not least, VA estimators need to be able to recover teacher quality on the basis of relatively few teacher-year observations, particularly as tenure decisions are often made quite soon after hiring.[1] The standard approach to this issue utilizes 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; Chetty et al., 2014a,b). To apply Bayesian shrinkage, papers estimating teacher VA have typically used the parametric empirical Bayes ('PEB') estimator, first proposed by James and Stein (1961). This estimator is appealing given its analytical convenience and the fact that it is the feasible version of the optimal Bayes rule for estimating teacher quality (under quadratic error loss) when unobserved quality is normally distributed.[2]

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 these two related issues.

The first contribution of our paper is to propose a feasible method for estimating teacher VA nonparametrically, without restricting attention to linear estimators. Following a standard setup in which residualized test scores equal underlying teacher quality (the heterogeneous teacher VA of interest) plus noise, we show that the teacher VA distribution can be identified nonparametrically, adapting a well-known result in the statistics literature to our context (Theorem 1 below).[3] Next, we present the nonparametric Bayes estimator for teacher VA (Theorem 2), drawing on a path-

[1] Among the US's 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.

[2] Further, even when unobserved quality does not follow a normal distribution, PEB is still the best *linear* estimator.

[3] A more general deconvolution proof of nonparametric identification is available in the literature, which we adapt to the case of teacher VA in Appendix B. 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. Then to assume that unobserved teacher quality in this formulation is also normally distributed involves an over-parameterization.

breaking 1956 paper by Herbert Robbins.[4] This estimator is optimal in the sense of minimizing the mean squared error of individual teacher VA estimates *regardless* of the true underlying VA distribution.

Comparing the parametric and nonparametric Bayes estimators, both can be written as functions of teacher fixed effects. Whereas the standard parametric Bayes estimator shrinks each teacher fixed effect linearly, the nonparametric Bayes estimator features a non-linear shrinkage rule, allowing the amount of shrinkage applied to each fixed effect to be non-monotonic. The latter estimator is infeasible, however, as it involves the unknown VA distribution. To obtain a feasible version, we apply a two-step approach (in Section 3). In essence, we first estimate the teacher VA distribution, denoted F, using nonparametric maximum likelihood, giving the estimated distribution $\hat{F}$ – the Bayesian prior estimated from the data.[5] Second, we plug this estimated distribution into the equation defining the nonparametric Bayes estimator from Theorem 2, given by the posterior mean. Doing so yields the nonparametric empirical 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). We leverage those advances in the current study, showing that the NPEB approach performs very well in Monte Carlo simulations in an environment that mimics typical administrative datasets in education. When the underlying distribution is normal, the performance of the NPEB estimator is almost indistinguishable from the PEB. When we consider more pronounced departures from normality, the NPEB approach continues to be highly responsive to the true VA distribution in a way that PEB is not. The simulations also shed light on the minimum sample sizes needed for the advantages of our nonparametric approach to become apparent. With samples of around four thousand teachers – far smaller than in many current education datasets – the NPEB estimator already gets very close to the infeasible 'oracle' estimator (see Section 4), noteworthy given NPEB makes no parametric assumptions about the underlying VA distribution.[6]

Next we apply the NPEB methodology using observational data, estimating teacher VA in two separate large-scale administrative datasets. 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. We find the estimated teacher VA distributions differ both from the normal distribution assumed in the bulk of the prior literature and from each other. In North Carolina, the estimated distribution has a relatively similar shape to the normal, although with more mass around zero and elongated tails,[7] while in the LAUSD, our estimated teacher distribution is skewed, with

[4]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.)

[5]The consistency of the nonparametric maximum likelihood estimator (or 'NPMLE') was established by Kiefer and Wolfowitz (1956); Lindsay (1995) provides an extensive treatment. We discuss the attractive properties of the NPMLE in Section 3 in the context of estimating teacher VA nonparametrically.

[6]The NPEB estimator has the additional attractive feature that it cures the selection bias problem which arises when considering tail observations – see Efron (2011). Measures of the degree of bias will provide a useful metric when comparing different candidate estimators in our simulations below.

[7]This aligns with Goldhaber and Startz (2017), who find that the teacher distribution in North Carolina is not

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 G).

The second main contribution of our paper is to explore the policy implications of these departures from normality. We do so for widely-discussed policies that release a given percentage of teachers, measuring the resulting policy gains in both the short and longer terms. Where departures from normality arise, our approach offers the potential to improve predictions of the policy gains (relative to PEB) – gains which could then be weighed by decision-makers against costs when formulating higher net-benefit policies. Our analysis will allow us to gauge the magnitudes of such predicted policy gains.[8]

We begin by considering the short-term gains of a proposal to release 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). Specifically, we compare the predicted test 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 significant differences in the estimated policy benefit, with PEB overstating test score gains of the policy by over 25 percent. To provide a fuller picture, we also simulate the short-run test score gains of policies that release *any* given percentage of teachers from the bottom of the VA distributions under the two approaches.

Alongside the impact of teacher VA on short-run test scores, we then compare the effects of our nonparametric VA estimates versus those under the parametric approach on long-run outcomes, including exit exam scores, drop-out rates, suspensions, and SAT scores. When considering the policy of releasing the bottom five percent of teachers according to VA, we find that PEB overestimates the long-run policy gains by around five percent in North Carolina; in the LAUSD, the overstatement of the long-run policy gains using PEB method is far larger, in the range 24-26 percent. The overestimation for the long-run outcomes in both datasets is very similar to our corresponding findings for short-run test scores.

Our methodology can be readily extended. Researchers may wish, for example, to incorporate classroom fixed effects or drift in the model. We show how to include classroom fixed effects, as

Gaussian, but the differences from the normal distribution tend to be small.

[8]While our focus is on policies to release teachers, we note that our framework is unsuited to the task of *ranking* teachers – the issue of whom to release (rather than what fraction to release). This is because the loss function used to derive the VA shrinkage estimator seeks to minimize VA estimation errors and not teacher ranking errors. In our context, if each teacher taught the same number of students, then teacher rankings under both empirical Bayes methodologies would be identical to those using simpler fixed effect methods (and each other). We will see below that the choice of estimator – whether fixed effect, parametric, or nonparametric empirical Bayes – has little appreciable impact on teacher rankings in either our simulations or empirical analyses. Gu and Koenker (2020) address the ranking issue in the context of nonparametric empirical Bayes explicitly.

they are quantitatively more important in our data.[9] Their inclusion is likely to reduce dispersion in teacher VA estimates, as class-level shocks are no longer attributed to the teacher. When we repeat the estimation and the policy analyses using the model with classroom fixed effects, doing so does reduce the variance in the VA estimates, as expected, 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 five-percent policy.

At a general level, our analysis underscores the plausible notion that the true unobserved distribution of teacher quality is likely to be context-specific. As a consequence, when the normality assumption is misplaced, we show how the performance of the shrinkage estimator can be improved by considering non-linear estimators in the form of NPEB. On the policy front, reforms whose implementation depends on teacher VA are influenced directly by the shape of the VA distribution, and hence the estimated policy gains when invoking normality may well differ significantly 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.

Our findings indicate that the differences for policy in some settings can be significant, compared with the parametric method. Further, as we show, the gains from the nonparametric approach become apparent even in quite moderate sample sizes. Thus, given the computational feasibility of our approach, analytical convenience need no longer weigh on the side of assuming normality. As reflected in other recent papers,[10] the nonparametric empirical Bayes methodology opens up new possibilities for empirical research; among these, it 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 our method in practice: Section 5 introduces the two administrative datasets, Section 6 describes the estimates of teacher VA using each dataset, and Section 7 presents the policy analysis, including short and longer term outcomes. Section 8 extends our method, and Section 9 concludes.

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 contributions of teachers) are additive in their effects. Accordingly, the achievement of a student i

[9] Given our data structure, it is not possible to include both classroom fixed effects and drift, as the teachers teach one class per year. To do that, one would need data in which teachers taught multiple classrooms per year.

[10] See, for example, Efron (2010), Dicker and Zhao (2016), Gu and Koenker (2017a), Gu and Koenker (2017b), Gu and Shen (2017), and Abadie and Kasy (2019), among others.

taught by teacher j in year t is written:

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

where $\tilde{y}_{ijt}$ is the student's observed test score (to be contrasted with y_{ijt} below, purged of covariates), and X_{ijt} captures 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.[11] The error term ϵ_{ijt} is assumed to be independently and identically distributed normal with variance σ^2_ϵ.[12]

The typical 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, given by the α_j's.

2.2 The Fixed Effect Estimator

Using the model in (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

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

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, and $n_j \equiv \sum_t n_{jt}$. Taking a weighted average of these sample averages $\{y_{jt}\}$ across all the classes taught by teacher j over time, with weights $h_{jt} \equiv n_{jt}/\sigma^2_\epsilon$, 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:

$$y_j \sim \mathcal{N}\big(\alpha_j, \sigma^2_\epsilon / n_j\big). \tag{2.4}$$

If the total sample $n_j \to \infty$ in the denominator of the expression for the variance, 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 the fixed effect estimator is a noisy estimator, especially for teachers beginning their careers. Shrinkage estimators (considered

[11] Rothstein (2017) and Chetty et al. (2017) discuss the validity of this assumption in the context of teacher value-added models. Our contributions to the value-added literature focus on the empirical Bayes procedure, rather than gauging bias in value-added measures due to potential non-random assignment.

[12] This normality assumption is made to simplify the following discussion. It is not necessary – see Appendix B.

next) account for this practical issue.

2.3 The Parametric Empirical Bayes Estimator

The current state-of-art estimator for teacher VA – the parametric empirical Bayes ('PEB') estimator, introduced first by Kane and Staiger (2008) and further developed by Chetty et al. (2014a) – leverages 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. The latter is the minimizer of the Bayes risk,[13] 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 PB estimator takes the following convenient form:

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

Several remarks about it are due:

1. When $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$, the posterior distribution of α_j conditional on y_j also follows a normal distribution, given by $\alpha_j | y_j \sim \mathcal{N}(y_j \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2/n_j}, \frac{\sigma_\alpha^2 \sigma_\epsilon^2/n_j}{\sigma_\alpha^2+\sigma_\epsilon^2/n_j})$, where the posterior mean of α_j, given the fixed effect y_j, is the optimal Bayes estimator of α_j.

2. The linear 'shrinkage' factor, $\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2/n_j}$, is always smaller than 1, which implies that the parametric Bayes estimator δ_j^{PB} shrinks the fixed effect estimator y_j toward zero.

3. The shrinkage factor is symmetric for fixed effects above and below zero (recalling that true teacher VA is centered on zero), being the same for all teachers with a given *total* sample size n_j (summing across all classrooms in all relevant time periods): the bigger the total sample size, the closer the shrinkage factor is to 1. 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.

4. The estimator δ_j^{PB} is infeasible since it involves the 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, obtained either through maximum likelihood or method of moments. (Appendix F provides details.)

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.[14] If this assumption is misplaced, the quality of

[13] This is defined as $\mathbb{E}[\sum_j (\hat{\alpha}_j - \alpha_j)^2]$, where j indexes teachers, α_j is the true VA of teacher j, and $\hat{\alpha}_j$ is an estimator for that.

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

the shrinkage estimator may deteriorate, perhaps significantly. It raises the further possibility that one might be able to find an alternative estimator that has a smaller Bayes risk; in what follows, we will seek to relax the normality assumption as well as enlarge the set of candidate shrinkage estimators to include nonlinear estimators.

2.4 The Nonparametric Bayes Estimator

Next we turn to our nonparametric Bayes estimator. As a precursor, the following theorem shows the distribution of teacher VA is nonparametrically identified, given the model in (2.2):

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.*

(See Appendix A for proof.)

This result implies that the data contain enough information about the distribution of true VA, and the normality assumption applied to unobserved teacher quality involves an over-parameterization; the over-parameterization also justifies why we are able to test whether normality holds in the data (see Appendix G). In the theorem, we assume a normally distributed error. Doing so is not necessary (as we show in Appendix B, using a result from a seminal paper by Kotlarski (1967)), although imposing normality on the noise term will be convenient for estimation purposes.

Next, we present the nonparametric Bayes ('NPB') estimator for teacher VA, using the teacher fixed effects $\{y_j\}$ and the model (2.4) as inputs, and defining the loss measure, $\mathcal{L}_2$ loss, as $L(\hat{\boldsymbol{\delta}}, \boldsymbol{\alpha}) \equiv \sum_{j=1}^{J} (\hat{\delta}_j - \alpha_j)^2$, where $\hat{\delta}_j$ is some estimator of true VA, α_j.

Theorem 2 *(Robbins, 1956) Given the model $y_j = \alpha_j + \nu_j$, with α_j independent of ν_j, and $\alpha_j \sim F$ and $\nu_j \sim \mathcal{N}(0, \sigma_\epsilon^2/n_j)$, then the estimator of α_j that minimizes Bayes risk under $\mathcal{L}_2$ loss 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 the expression

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

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

(See Appendix A for proof.)

Theorem 2 presents two expressions for the nonparametric Bayes estimator. The expression in (2.6) is simply Bayes rule, which takes the form of the posterior mean of α_j, with the distribution

F playing the role of the prior distribution of α_j in the Bayesian framework. The expression in (2.7) is known as "Tweedie's formula" in the literature (see Robbins (1956) and Efron (2011)), and makes the estimator's nonlinear shrinkage explicit.

Several remarks about the nonparametric Bayes estimator 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.[15]

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

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 Tweedie's formula introduces a non-linearity into the shrinkage rule with respect to y_j.

4. Tweedie's formula does not automatically 'shrink' the fixed effect estimator towards zero; unlike parametric Bayes, shrinkage can be non-monotonic. Both the direction and magnitude of the shrinkage are likely to be context-specific.

In light of our teacher focus, we note that teachers with the highest variances will not necessarily be shrunk the most toward zero under Tweedie's formula, a feature that turns out to be empirically relevant. Suppose a new teacher j with a small total sample size n_j relative to σ_ϵ^2 happens to have a high y_j – a teacher who performs very promisingly in her early years in the school system, for example. Under the parametric shrinkage rule, this teacher will be heavily discounted, with her VA measure being shrunk significantly toward zero.[16] 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.[17]

Example: We illustrate ways in which shrinkage may operate under the nonparametric Bayes estimator compared with parametric Bayes.[18] To that end, we assume true teacher VA has distribution F, which takes the specific form

$$F = 0.95\mathcal{N}(0,\theta_1) + 0.025\mathcal{N}(-1,\theta_2) + 0.025\mathcal{N}(1,\theta_3)\,. \tag{2.8}$$

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

[16]This is simply because the normal distribution deems it unlikely that this teacher has a high VA, given the tails must be thin, the large fixed effect arising instead purely by chance due to the associated high variance.

[17]A similar thought experiment can be conducted for a teacher whose observed performance is very poor (that is, who has a large negative y_j) with 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, rather allowing for a range of possibilities that are informed by the data.

[18]Because we use a *known* distribution in the example, this is a comparison between NPB and PB estimators.

This mixed normal distribution has the built-in feature that the tails have small probability masses concentrated around the values -1 and 1 respectively, while the majority of the mass follows a normal distribution centered at zero.[19] Artificiality to one side, this example gives insight into the way deviations from normality in the tails can lead to nonlinear, and even non-monotonic, shrinkage behavior.

Figure 1 compares the amount of shrinkage for the parametric and nonparametric Bayes estimators, respectively, when we set $\theta_1 = \theta_2 = \theta_3 = 0.03$, which implies that the variance of the true VA is 0.08; hence we use $\sigma_\alpha^2 = 0.08$ in the parametric Bayes estimator for comparison. The figure does so for a hypothetical teacher j with total sample size $n_j = 16$ whose fixed effect estimate, y_j, is in the range of $[-1.5, 1.5]$.

As the figure illustrates, the parametric Bayes estimator (shown by the diagonal straight line) shrinks the fixed effect estimates linearly and symmetrically toward the mean of zero, with the direction of the shrinkage depending only on the sign of y and not its magnitude. In contrast, the NPB estimator (in equation (2.7)) is in general a nonlinear function of y and the direction of shrinkage depends not only on the sign of y but also its magnitude. Differences in the amount of shrinkage between the two estimators are especially pronounced at the mass points of -1 and 1. The parametric Bayes estimator does not account for the presence of these mass points – they are assumed away by normality – and so shrinks them toward zero. The NPB estimator, in contrast, adapts to the underlying distribution and pulls fixed effects nearby (below and above) to the two mass points in the tails. In doing so, it accounts for the non-negligible probability that the true quality of some teachers may take values in the vicinity of these mass points.

3 A Feasible Nonparametric Bayes Estimator

The nonparametric Bayes estimator just presented, δ_j^{NPB}, is infeasible in practice since it involves unknown quantities – specifically, σ_ϵ^2 and the distribution F. This section develops a feasible version, which we refer to as the nonparametric empirical Bayes (or NPEB) estimator. In the same spirit as the parametric empirical Bayes method, where the unknown parameters are estimated from the data directly, we will recover the unknown parameter σ_ϵ^2 using its maximum likelihood estimator (see Appendix F), and estimate the distribution F using nonparametric maximum likelihood rather than assuming it belongs to a parametric distribution family.

3.1 Nonparametric Maximum Likelihood Estimation of the Distribution F

In order to implement the Bayes estimator in Theorem 2 in practice, we must somehow learn the 'prior' distribution F. A fundamental insight due to Robbins is that it is both desirable and feasible to estimate F nonparametrically.[20]

[19] Here, we set $\sigma_\epsilon^2 = 0.25$, which is roughly the same as in the North Carolina mathematics score data.
[20] See the abstract of Robbins (1950).

On the desirability front, Kiefer and Wolfowitz (1956) established the consistency of the nonparametric maximum likelihood estimator ('NPMLE') for the mixing distribution F. The NPMLE in mixture models has been found to have attractive properties, despite optimizing over an infinite-dimensional parameter space, as results in the recent statistics literature indicate, notably by Zhang (2009), Saha and Guntuboyina (2020), and Polyanskiy and Wu (2020).[21]

On the feasibility front, applications of the nonparametric method have been facilitated by recent computational advances. Here, we use methods proposed in Koenker and Mizera (2014) and Koenker and Gu (2017) in order to recover the teacher VA distribution. Those papers, building on Robbins (1956), 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. The methodology turns out to be suited to many contemporary 'Big Data' applications,[22] with the teacher VA application fitting well within this general framework. Here, teacher quality can be thought of as unobserved heterogeneity in a model of test scores that accounts for variation unexplained after controlling for all observed heterogeneity captured by X_{ijt}.

To apply the framework, we denote the distribution of α_j 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 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 NPMLE for F.

Until recently, the standard approach to computing the NPMLE involved the EM algorithm due to Laird (1978),[23] although the algorithm's slow convergence in nonparametric empirical Bayes applications has inhibited the NPMLE's widespread implementation. Against that backdrop, Koenker and Mizera (2014) proposed an alternative, and much faster, computational method for the NPMLE. They showed 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

[21] The statistics literature has also established the consistency of the marginal density g in the Tweedie formula in terms of Hellinger distance (see van de Geer (1993) for the homogeneous variances case and Jiang (2020) for the heterogeneous variances case). Saha and Guntuboyina (2020) have shown the near-parametric risk of this marginal density when the distribution is F with a finite number of support points, implying the noteworthy adaptive property of the NPMLE. Recent work by Polyanskiy and Wu (2020) discusses the self-regularization property of NPMLE when F has sub-Gaussian tails, suggesting that the NPMLE of F in a mixture model already automatically selects a parsimonious estimator without penalization, a property not expected for the usual nonparametric maximum likelihood estimator; the mixture model structure acts as a shape restriction in this context.

[22] See Efron (2010) for a survey. Intuitively, a compound decision framework arises naturally in settings with large amounts of data under the modeling assumption that the latent variable is drawn from a common distribution. In an environment with many related decision problems that involve learning individual α_j's, using everyone's data to uncover the common distribution F can be shown to improve collective performance.

[23] See Heckman and Singer (1984) for an early econometric application.

much wider range of applications of the nonparametric method for models with heterogeneity.[24] Here, the large-scale data involved in teacher VA estimation make such applications likely to benefit from the scalability of the new computational method. (Appendix H discusses computational issues further.)

3.2 The Plug-in Nonparametric Empirical Bayes Estimator for VA

With the NPMLE $\hat{F}$ in hand, we construct the NPEB VA shrinkage 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 in equation (3.2) is the feasible version of the posterior mean defined in equation (2.6) in Theorem 2.[25] Having obtained the NPMLE for F based on (3.1), evaluating (3.2) only involves matrix operations, as $\hat{F}$ takes a discrete form – see Appendix H.

The NPEB estimator has been shown to have desirable statistical properties, despite the NPMLE of F having a slow convergence rate (Fan, 1991). In particular, comparing the risk of the NPEB estimator with the infeasible NPB estimator, the difference tends to zero asymptotically when F has compact support, thus justifying the asymptotic risk optimality of the NPEB – see Jiang and Zhang (2009) and Saha and Guntuboyina (2020) for the homogeneous variance case (e.g., all teachers have the same class size) and Jiang (2020) for the heterogeneous variance case. Given these theoretical results, our proposed estimator δ_j^{NPEB} has the potential to improve on the linear PEB estimator in the sense of having a smaller risk, and perhaps significantly so, when the underlying distribution F cannot be approximated well by the normal distribution. The magnitude of this improvement can be evaluated through simulations, which we turn to next.

4 Simulations

In this section, we compare the performance of three candidate estimators – the fixed effect estimator, the PEB estimator, and our NPEB estimator – using illustrative simulations in which the distribution F used to generate the data is known. The comparisons are relative to a benchmark, infeasible in practice, in which the researcher knows the true underlying teacher quality distribution and is therefore able to use the optimal Bayes rule to construct the oracle estimator.

We generate the data based on the model $y_j = \alpha_j + \epsilon_j$ specified in (2.4), where ϵ_j follows a normal distribution with mean zero and variance σ_ϵ^2/n_j, and n_j is the class size of individual teacher j.[26] In the simulations, we start by fixing the total number of teachers (J) to be 10,000 – we will

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

[25] We discuss the construction of the estimator in more detail in Appendix H. There, we also explain why the feasible version of the NPB estimator is constructed using (2.6) directly, rather than equivalent expression given in Tweedie's formula.

[26] We will think of the teacher as having just one year of experience (having taught only one class), in which case her VA estimate can differ markedly from her fixed effect estimate.

vary J below when examining finite sample properties – and we set $\sigma_\epsilon^2 = 0.25$ to mimic the actual estimates we will obtain using the North Carolina and LA data. Two alternatives are compared: a homogeneous class size case where every teacher has a class size of 16, and a heterogeneous class size case where class size is drawn randomly from the set $\{8, 16\}$ with equal probability.[27]

The simulation sample to hand, we implement the three candidate estimators. In particular, we estimate the variance of α under PEB by maximum likelihood under the assumption that F follows a normal distribution, and use this to construct the linear shrinkage estimator. For NPEB, we estimate the VA distribution F using the nonparametric maximum likelihood method, and plug this in to (3.2). The fixed effect estimates are the $\{y_j\}$ themselves.

The performance of the three estimators is then assessed under four different known distributions. The first two are the normal and mixed normal, both symmetric around zero. We also consider two asymmetric VA distributions: the skewed normal and the skewed logistic, which has a thinner tail than the normal, both distributions being calibrated to have means of zero.

We use two performance measures to compare the estimators. The first is the sum of squared errors ('SSE'), the empirical counterpart of risk, which takes the form $\sum_{j=1}^{J}(\delta_j - \alpha_j)^2$, where α_j is the true VA for teacher j, δ_j is a VA estimator for teacher j, and J is the sample size (total number of teachers) in the simulations. The second is a measure of selection bias in the bottom 5 percent, estimated as $\frac{1}{0.05J}\sum_j(\delta_j - \alpha_j)1\{\delta_j < t\}$, with t being the 5% quantile of the fixed effects. To motivate this second measure, the fixed effect estimator is known to suffer from selection bias in the sense that teachers with more extreme (larger or smaller) fixed effect estimates tend to have less extreme true VA. Efron (2011) noted that the Bayesian shrinkage estimator is capable of correcting such selection bias, our bias measure indicating the extent to which NPEB has an advantage over other estimators in this regard.

Teacher Quality Distribution is Normal: Table 1(a) displays the simulation results when teacher quality is normally distributed according to $F \sim \mathcal{N}(0, 0.08)$. Here, the normality assumption built into the PEB estimator is correct, so it performs almost identically (given the variance of the normal distribution is still estimated) to the infeasible oracle estimator, which uses the true distribution. We see that the PEB estimator improves on the sum of squared errors and the left tail bias of the fixed effect estimator substantially. At the same time, it only outperforms our NPEB estimator by a very small margin: whether considering the homogenenous or heterogeneous class size cases, the risk ratios (NPEB/PEB) are very close to 1 (131.1/130.5, and 187.0/186.3), and the left tail bias is identical, a striking result given NPEB does not make any parametric assumption about the distribution F.

Teacher Quality Distribution is Non-Normal: Next, Tables 1(b), (c), and (d) show simulation results when teacher quality is not normally distributed. Specifically, in Table 1(b), true

[27] Under homogeneous class sizes, the choice of estimator has *no* effect on teacher rankings since both PEB and NPEB are monotone transformations of y_i. With heterogeneous class sizes, we no longer have monotonicity, although all estimators have very similar misclassification rates for the four VA distributions we have considered – see Appendix Tables I.2(a)-I.2(d).

teacher quality follows the mixed normal distribution $F \sim 0.95\mathcal{N}(0, 0.03) + 0.025\mathcal{N}(-1, 0.03) + 0.025\mathcal{N}(1, 0.03)$ as in equation (2.8),[28] while asymmetric non-normal distributions are shown in Tables 1(c) and 1(d): in Table 1(c), true teacher quality follows the skewed normal, and in 1(d), a skewed logistic distribution.[29]

It is noteworthy that our NPEB estimator achieves a sum of squared errors very close to that of the infeasible estimator (which assumes the true distribution is known) for all four VA distributions, whether considering homogenenous or heterogeneous class size cases. For the three non-normal distributions, the starkest contrast between the performance of the NPEB and PEB in terms of overall risk (under squared error loss) occurs when true VA follows a mixed normal distribution: the SSE under NPEB is far smaller than under PEB, with the performance advantage still apparent for the other two distributions. The table also shows the NPEB corrects the selection bias of the fixed effect estimator, displaying no selection bias for the bottom 5 percent for all three non-normal distributions. Further, the PEB is outperformed by NPEB (in terms of selection bias) in every panel in the table when normality is misplaced.

Finite Sample Properties: Since the NPEB estimator relies on the nonparametric maximum likelihood estimator of the VA distribution while the fixed effect and PEB estimators are both parametric, it is important to investigate the NPEB's finite sample performance. To that end, we repeat the previous simulations, considering successively larger samples of teachers, drawn in turn from the set $\{100, 500, 1000, 2000, 4000, 10000, 12000\}$, many administrative datasets being in the ballpark of the largest value in the set.

We plot the results for PEB and NPEB estimators along with the infeasible Bayes estimator in Figure 2 for all four VA distributions used in the simulations.[30] As sample size increases, the risk of our NPEB estimator converges to the infeasible benchmark quickly, approaching a risk ratio of 1 when the number of teachers equals a thousand; when the number of teachers exceeds four thousand, our nonparametric estimator is near-identical to the infeasible estimator in terms of SSE. In stark contrast, the PEB estimator does *not* improve as sample size increases, since the estimated VA distribution is not adaptive.

When the true VA distribution is normal, the performance of the nonparametric estimator based on mean squared error becomes virtually identical to PEB once the number of teachers exceeds 4,000. When the true VA distribution is non-normal, the NPEB estimator outperforms the PEB estimator in far smaller samples – no more than 500 teachers across all three cases.

[28] For comparability, the mixed normal distribution has the same first three moments (mean, variance, and skewness) as the normal distribution in Table 1(a).

[29] The skewed normal still has a normal tail while the distribution is skewed to the right with location parameter -0.4 and shape parameter 5; the scale parameter is calibrated so that the mean of F equals zero. The skewed logistic has a thinner tail than the normal and is also skewed to the right. The location and shape parameters are set to be -0.5 and 5 respectively, and the scale parameter is also calibrated so that the mean of the distribution is zero. The variance of the resulting true VA is roughly 0.1 in both cases.

[30] We only plot the results for heterogeneous class sizes (to conserve space), although results with homogeneous class sizes look very similar. The fixed effect estimator is omitted in these figures since it performs so much worse than the parametric or nonparametric empirical Bayes estimators in terms of mean squared error and tail bias (as shown in Tables 1(a) - 1(d)).

Next, we report the selection bias for different sample sizes, comparing the estimated VA using NPEB and PEB and the true VA for the lower 5 percent of the sample (see Figure I.1). The results show our NPEB estimator essentially cures all selection bias, now even with small samples, tracking the infeasible estimator very closely indeed. In contrast, the PEB estimator exhibits substantial selection bias when there are departures from normality, underlining that the linear shrinkage estimator does not eliminate selection bias when the VA distribution is misspecified.

In sum, the simulations indicate that our approach has an appealing versatility. For the typical sample sizes used in VA applications, the proposed NPEB estimator performs similarly to the PEB estimator when the true distribution is normal, despite being a nonparametric method. When the true distribution is not normal, our approach outperforms the PEB estimator by a substantial margin, performing almost as if the true distribution were known for moderate sample sizes of 4,000 or more. 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 – a powerful attribute when taken to observational data, which is our focus in the following sections.

5 Data

Our data are drawn from two administrative datasets, each providing detailed information about students and teachers. While covering similar time periods, grades, and demographic information, we discuss each data source separately given there remain important differences between them; a more detailed description is provided in Appendix C.

North Carolina: Our first administrative dataset[31] covers all public school students in North Carolina for third through fifth grade – third grade from 1996-97 to 2008-09 and fourth and fifth grades from 1996-97 to 2010-11.[32] In total, the dataset consists of around 1.85 million students with 4.5 million student-year observations. It also provides 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 in constructing the sample used to estimate 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 making sample restrictions (described in Appendix C), our final sample consists of approximately 2.7 million student-year observations, covering 1.4 million students and 35,000 teachers.

[31]The relevant data citation is: North Carolina Education Research Data Center (1996-2017).

[32]Our analysis is restricted to students in grades three to five since our dataset records the test proctor rather than the class teacher explicitly, and the test proctor is typically the teacher who taught the students throughout the year in these grades. Data for third grade ceases following 2008-09 because the third grade pretest was discontinued after that year.

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 dataset. 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.

Los Angeles Unified School District: Our second data source comes from the Los Angeles Unified School District (LAUSD).[33] The dataset 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,[34] and 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, a grade repetition indicator, and parental education (missing for thirty percent of sample). Similar to the North Carolina dataset, 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 provide summary statistics for the LAUSD data, column (3) for the entire sample, and column (4) for the VA analysis dataset. Similar to North Carolina, 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 datasets, clear differences in samples are apparent. North Carolina has a majority-white student body with a large black minority, whereas the LAUSD is majority-Hispanic. The LAUSD sample is also drawn from students with significantly more disadvantaged backgrounds, 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 affect the underlying distributions of teachers found in these two settings.

6 Results

This section reports estimates of teacher VA using our proposed NPEB methodology, alongside estimates using the PEB approach. We describe results for North Carolina and the LAUSD in turn. Then we discuss how these differences are relevant for teacher rankings, and compare the out-of-sample performance of our NPEB estimator relative to the PEB estimator.

[33] Data citation: Los Angeles Unified School District (2003-2017).

[34] 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.

6.1 VA Estimates

North Carolina: Figure 3(a) provides a boxplot of teacher *fixed effects* in our North Carolina dataset for teachers who appear once, twice, three times, and more than three times, respectively. It shows that teachers appearing for more periods – typically more experienced teachers – exhibit less dispersion as their fixed effect 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, the boxplot in the lower panel (Figure 3(c)) showing the magnitude of the *shrinkage* applied by our NPEB estimator. As expected, teachers with more than three years of data receive only small amounts of shrinkage, while teachers who appear less frequently are shrunk toward zero in a more pronounced way.

Figure 4(a) shows our nonparametric estimate of the teacher VA distribution with the solid line, estimated using equation (3.1) – we plot the relevant quantiles for visual ease. We also estimate the distribution of teacher VA in North Carolina under the normality assumption using maximum likelihood,[35] finding that teacher VA is distributed $\mathcal{N}(0, 0.047)$, which implies a standard deviation of 0.217.[36] We superimpose the (assumed) normal quantiles with the dashed line in Figure 4(a) so that deviations from normality can be visualized.

The figure makes clear that NPMLE yields a VA distribution that has more mass around zero and less mass for intermediate VA values compared to the Gaussian distribution. In terms of the tails, differences between the NPMLE and Gaussian density estimates are almost negligible in North Carolina, aside from rather subtle differences.[37]

Next we consider *shrinkage*. Differences between the NPMLE and Gaussian estimates of the VA distribution give rise to distinct shrinkage rules, leading to different NPEB and PEB estimates of a given teacher's quality. Panels (a) and (b) of Figure 5 plot estimated teacher VA in North Carolina using our NPEB estimator against the PEB estimator in the policy-relevant bottom and top tails of the distribution, respectively.

Specifically, we plot those teachers whose fixed effect estimates are below the 5^{th} (on the left) or above the 95^{th} percentile of the distribution (on the right), with the vertical dashed lines demarking the bottom and top one percent of observations (on the left and right, respectively). In panel (a), focusing on the far left tail, we see that teachers with their PEB estimates below -0.5 (these being the very bottom 0.5% of teachers in terms of their FE) are situated beneath the 45-degree line, indicating that the NPEB estimator shrinks these teachers less toward the mean of zero compared to PEB. This is due to the fact that the NPMLE finds slightly more mass in the extreme left tail, and so shrinks these teachers less as it expects more teachers to have true VA there compare to

[35] 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.

[36] The other input to the PEB shrinkage estimator is estimated to be $\hat{\sigma}_\epsilon^2 = 0.248$.

[37] NPMLE finds slightly more mass in the extreme left tail (VA<-0.6, which corresponds to the bottom 0.3% of teachers) compared with the Gaussian, but slightly less mass in the rest of the left tail above -0.5 (which demarks the bottom 1% of teachers). In the very right tail, NPMLE detects slightly more mass in the top 1% of teachers.

the Gaussian. For the right tail, the estimates only disagree in the far right, above 0.6 (accounting for the very top 0.1% of teachers), with the NPEB shrinking these teachers less toward the mean given NPMLE found more teachers in the extreme right tail than the Gaussian.

LAUSD: Following the same results progression as for North Carolina, Figures 3(b) and 3(d) present boxplot estimates for LAUSD – teacher fixed effects and the magnitude of shrinkage that our NPEB estimator applies based on 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 NPMLE is shown in Figure 4(b), using the solid line. Based on the PEB estimator, we find teacher VA is distributed as $\mathcal{N}(0, 0.0977)$, implying a standard deviation of 0.3126, which is considerably higher than the estimated variance in North Carolina.[38] This normal distribution is superimposed using the dotted line in Figure 4(b). What stands out visually is the distinct lack of symmetry: the solid line representing the NPMLE quantile lies above the dotted line indicating the Gaussian quantile in the left third of the distribution but not on the right. Clearly, the nonparametric VA distribution is skewed, with a much thinner left than right tail.

Turning next to shrinkage, the thinner left tail gives rise to large differences between NPEB and PEB estimates of a given teacher's VA, as shown in the bottom two panels of Figure 5. These plot estimated teacher VA from our NPEB against the PEB estimator for teachers whose fixed effect estimates are below the 5^{th} (on the left) or above the 95^{th} percentile of the distribution (on the right). For teachers in the left tail, their NPEB estimates are larger than their PEB estimates (being located above the 45-degree line), indicating that the NPEB shrinks these teachers further toward the mean.

This shrinkage behavior is driven by the thin left tail estimated using NPMLE; the NPEB shrinks left-tail teachers more toward the mean since the NPMLE finds a smaller mass of teachers here compared to the Gaussian. In contrast, the PEB and NPEB estimates largely agree for teachers in the right tail since the NPMLE and Gaussian estimates of the right tail of the VA distribution are relatively similar. Our nonparametric method therefore adapts to the skewness of the distribution by shrinking teachers with fixed effect estimates below the 5^{th} percentile far more strongly back toward the mean, while applying similar shrinkage to the PEB when the fixed effect estimates take values above the 95^{th} percentile.

Teacher Rankings: It is natural to ask 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. (In this context, a Type I error occurs when a teacher is ranked below 5% while her true quality ranking is above 5% – conversely, for a Type II error.) With homogeneous class sizes, the PEB and NPEB estimators and the fixed effect estimators will all produce the same teacher rankings and thus misclassification rates (Guarino et al., 2015; Bitler et al., 2019). Once

[38] The other parameter is estimated as $\hat{\sigma}_\epsilon^2 = 0.2596$.

we allow class size to be different, however, the PEB 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 chosen, in turn making it unclear which estimator should (in principle) be used to determine *whom* to replace.[39]

In practice, we find that the choice of empirical Bayes method has little appreciable impact on teacher rankings. In both datasets, a very small fraction of teachers ranked in the bottom five percent of the teacher quality distribution under PEB are *not* ranked in the bottom five percent according to NPEB. (The exact numbers 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 I.2(a)-I.2(d)).

6.2 Out-of-Sample Predictions

Given our NPEB estimates of VA, we now evaluate the performance of our NPEB estimator relative to the fixed effect and PEB estimators on the basis of their ability to predict future outcomes. Suppose, for instance, that school boards or policymakers observe a teacher's past performance and wish to predict future outcomes for that teacher. We can measure predictive performance 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 performance, starting when the teacher first appeared in the sample up until the t^{th} period.

This prediction exercise faces a difficulty in that the class sizes of teachers in period $t+1$ likely 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 is, 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 'large class size' teacher. To account for this, we use the following two measures of prediction accuracy proposed by 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.[40] In turn, TMSE is given by:

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

[39] The appropriate loss function when determining whom to replace should quantify misclassification as the criterion for the optimal estimator, instead of using the sum of squared errors of the true and estimated VA. Gu and Koenker (2020) discuss suitable alternative loss functions for the purposes of ranking and selection.

[40] The NMSE is the usual sum of squared errors with an adjustment term $n_{j,t+1}$ (which serves to scale back the contribution of teachers with large classes sizes, who have higher precision by construction, by the size of their class, $n_{j,t+1}$).

Without the adjustment term $\frac{\sigma_\epsilon^2}{n_{j,t+1}}$, the quantity is just the usual sum of squared errors, the adjustment being 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, respectively, for the NPEB, PEB, and fixed effect estimators. Each row in the table represents the number of prior years of teacher j's performance that is used to make the prediction.[41] The two empirical Bayes methods (reported in the first two columns) substantially outperform the fixed effect method when only a few years of prior data are used, the extra 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.

Now 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 dataset. When 2-5 years of data are used, the predictive 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 PEB, 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 shrunk 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 (as stated in the Introduction), namely using 2-5 years of data per teacher.

7 Policy Analysis

The discussion in Section 2 made clear that empirical Bayes VA estimates take the form of posterior means of the VA distribution, applying shrinkage to the teacher fixed effects they are associated with. Estimates in the previous section indicate that the amount of shrinkage depends on the shape of the VA distribution, as well as the magnitude of the fixed effects. Now we examine how differences in the VA distribution can affect education policy calculations when comparing parametric and nonparametric empirical Bayes methodologies.

We focus on widely discussed policies that target the bottom of the teacher quality distribution (noting one could also consider the top of the distribution). When evaluating such policies, the measures we consider involve average gains from replacing teachers at the bottom of the VA distribution by average-quality teachers – policy gains that depend on the VA distribution obtained under the parametric and nonparametric approaches. Our interest is in gauging the extent to which average policy gains are over- or under-stated, comparing the methods.

Evidence using large-scale administrative datasets in the previous section shows that disagreement between the VA distribution (relative to the normal) in both tails of the North Carolina distribution and in the right tail of the LAUSD distribution are relatively minor. Thus, policies

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

targeting the tails of the NC distribution or the right tail of the LAUSD distribution would be expected to be relatively invariant to the choice of nonparametric versus parametric estimator. At the same time, the normality assumption is particularly misspecified in the left tail for the LAUSD, which is important for policies targeting low-VA teachers – teacher release policies, for example. Here, the PEB methodology (taking the VA distribution to be normal) is likely to *overstate* the benefits of such policies since the nonparametric estimator finds less relevant mass in the left tail.

We examine a policy that has gained considerable traction, seeking to replace poor-quality teachers. The specific proposal made by Hanushek (2009, 2011) and further explored by Chetty et al. (2014b) involves releasing teachers in the bottom 5% of the estimated VA distribution and replacing them with average-quality teachers. We consider a more general policy that replaces the bottom q% of teachers, while paying particular attention to the 'bottom 5%' cutoff, given its prominence in both the prior literature and the policy debate.[42]

In what follows, we first quantify how features of the VA estimates from North Carolina and the LAUSD affect the short-run benefits of such teacher-release policies under the nonparametric versus parametric methodologies measured in terms of test score gains. Then we consider the effects of teacher-release policies in terms of longer-run outcomes in both settings.

7.1 Short-run Effects

We will assess the short-run effects of teacher-release policies in terms of the average test score gains among students whose teachers were replaced, which depend on the 'input' VA distribution under nonparametric and parametric approaches. These test score gains are a function of the cutoff percentile q, noting that the VA distribution is centered to have mean zero, and a one-unit increase in VA (or α) leads to a one-SD test score gain.[43]

The average test score gain is defined formally in the context of teacher-release policies by:

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

given the cutoff percentile q, where $F^{-1}(q)$ denotes the cutoff value of α that designates the q^{th} percentile for (general) VA distribution $F(\alpha)$. (Thus, the proportion $q = \int_{-\infty}^{F^{-1}(q)} dF(\alpha)$, for reference.)

We will index the average test score gains (via superscripts) according to the approach used to obtain the VA distribution. Using equation (7.1), the average gain under NPEB can be written:

$$G^{NPEB}(q) = -\mathbb{E}_F[\alpha \mid \alpha < F^{-1}(q)] = -\frac{\int_{-\infty}^{F^{-1}(q)} \alpha dF(\alpha)}{\int_{-\infty}^{F^{-1}(q)} dF(\alpha)}\,, \tag{7.2}$$

assuming general VA distribution F. Under PEB, assuming that teacher VA is normally distributed

[42] Results from a corresponding analysis of teacher-retention policies are available on request.

[43] The gains can be transformed into the average test score gain among all students, multiplying the average by the cutoff percentile q.

with $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, the average gain is given by

$$G^{PEB}(q) = -\mathbb{E}_\Phi[\alpha \mid \alpha < \Phi^{-1}(q)] = -\frac{\int_{-\infty}^{\Phi^{-1}(q)} \alpha d\Phi(\alpha)}{\int_{-\infty}^{\Phi^{-1}(q)} d\Phi(\alpha)}, \tag{7.3}$$

where Φ is the cdf of $\mathcal{N}(0, \sigma_\alpha^2)$.

In practice, neither the key parameter in the normal distribution (σ_α^2) assumed under PEB nor the distribution F are known. Thus, we have to estimate both objects under the two approaches: $\hat{F}$ in (7.2) using NPMLE and $\hat{\sigma}_\alpha^2$ in (7.3) using MLE. An additional practical difficulty when implementing the policy concerns identifying the bottom q teachers, as the decision rule relies on true VA being observed, which makes it infeasible. In practice, we replace α with a VA estimate and construct the cut-off value based on its empirical quantile.[44] The VA estimates are then obtained using three years of simulated data for each teacher and assuming that teachers all have class sizes of twenty, calculating the sample analogs of both test score gain measures for the two distributions via Monte Carlo simulation.[45] The sample analog gain under $\alpha \sim \hat{F}$ becomes:

$$\widehat{G}^{NPEB}(q) = -\mathbb{E}_{\hat{F}}[\alpha | \hat{\alpha} < \hat{Q}_q(\hat{\alpha})] .$$

Correspondingly, the sample analog under $\alpha \sim \mathcal{N}(0, \hat{\sigma}_\alpha^2)$ is:

$$\widehat{G}^{PEB}(q) = -\mathbb{E}_{\mathcal{N}(0,\hat{\sigma}_\alpha^2)}[\alpha | \hat{\alpha} < \hat{Q}_q(\hat{\alpha})] ,$$

where in both equations, $\hat{Q}_q(\hat{\alpha})$ is the empirical q-quantile of the estimated $\hat{\alpha}$.

The results based on these simulations are plotted in Figure 6. They do so separately for North Carolina and the LAUSD, for different percentages of teachers being released (given on the x-axis in each panel). Several features are apparent: First, the policy gains are lower when using estimated rather than true VA, since some teachers with true VA below the given cutoff percentile are retained. The decreases in test score gains are relatively modest, however, and are similar for the PEB and NPEB methodologies in the two panels, as indicated by the downward shifts in the dashed (relative to the solid) lines.[46] Second, the gains decline steeply using either method in both datasets when the lowest-VA teachers are replaced, with the profiles flattening out as the replacement cutoff rises up the VA distribution (as expected). Third, while using the PEB method gives rise to a slight

[44] Essentially, our practical implementation of the policy is to construct estimates of teacher VA, rank them and select the bottom q proportion to be released.

[45] The relevant steps are as follows: Under the PEB methodology, we first sample 40,000 observations from $\mathcal{N}(0, \hat{\sigma}_\alpha^2)$. 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 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 $\frac{1}{40000q} \sum_j \alpha_j 1\{\delta_j^{PEB} \leq \hat{Q}_q(\delta^{PEB})\}$ as an estimator for $\widehat{G}^{PEB}(q)$, the sample analog of equation (7.3). By the law of large numbers, this produces a consistent estimator for $\widehat{G}^{PEB}(q)$. Analogously, we sample VA from the distribution $\hat{F}$, use the NPEB method to obtain an estimator δ_j^{NPEB}, and calculate $\widehat{G}^{NPEB}(q)$ in a similar fashion.

[46] 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.

overstating of the gains in North Carolina relative to NPEB (see panel (a), comparing the two dashed lines), the overstating of gains in the LAUSD dataset is far more pronounced, as shown in panel (b).

Table 4 reports the policy gains from the teacher-release policy when true teacher VA is unobserved: these are the numbers underlying the dashed lines in Figure 6. Taking the bolded row corresponding to releasing the bottom 5 percent of teachers and replacing them with those of average quality, the PEB method overstates the policy gains by seven percent in North Carolina and by fully twenty-six percent in the LAUSD data; results are very similar to the case when VA is observed by the policymaker – see Table I.4. Thus, our findings indicate that the estimated short-run policy gains for teacher-release policies are somewhat overstated in North Carolina and significantly overstated in LAUSD under the parametric method. Our nonparametric methodology allows possible discrepancies to be assessed on a context-by-context basis in these settings and elsewhere.

7.2 Implications for Long-Run Outcomes

Alongside the effects of teacher VA on short-run test scores, we now consider long-run outcomes. Our data allow us to examine several important measures, including: drop-out rates, suspensions, SAT scores, and exit exam scores (see Appendix C.3 for a detailed data description). Using those, we focus on quantifying the benefits under the benchmark policy that replaces the bottom five percent of teachers according to VA that has been prominent in the recent literature.

The analysis requires that we link the long-run outcomes at our disposal with teacher VA. Here we follow the method proposed by Chetty et al. (2014b), calculating the long-run benefits of being assigned to a higher-VA teacher for one grade.[47] These take the form of slope coefficients associated with PEB and NPEB, respectively, which feed into the calculations.

Tables 5(a) and 5(b) report the policy gains in terms of long-run outcomes from releasing the bottom five percent of teachers according to VA, for North Carolina and for LAUSD. These long-run gains mimic the gains in the test score results above, representing gains among students whose teachers are replaced. Calculations using the parametric and nonparametric method are given in Panels A and B of each table. The policy gains in the third row of each of the four panels are determined by multiplying the increase in the long-run outcome resulting from having a teacher one standard deviation higher in the distribution, $\hat{\kappa}$, in the first row by the average change in VA as a consequence of releasing bottom 5 percent teachers, $\hat{\Delta}m_\sigma$, in the second row.[48] The last row of each table then indicates the degree of overestimation when imposing normality.

Comparing the first rows of Panels A and B in Table 5(a) makes clear the slope coefficients

[47] This method is familiar: for completeness, we offer a brief description in Appendix D.

[48] The specific formula for the long-run outcome gains under PEB and NPEB is given by:

$$G^d = \hat{\Delta}m_\sigma^d \times \kappa^d, \;\; d \in \{PEB, NPEB\}, \tag{7.4}$$

where $\hat{\Delta}m_\sigma^d$ represents the estimated average improvement in VA (measured in terms of test scores) of the policy assuming that true VA is unobserved by the policymaker.

when applying the two approaches are similar in the North Carolina data. In contrast, Table 5(b) shows that a one-SD increase in VA yields greater long-run effects under PEB compared to NPEB in the LAUSD data.[49]

In terms of the overall long-run gains using the two methods, we find that PEB overestimates policy gains in terms of long-run outcomes in a tight band, between four and five percent, in North Carolina, depending on the outcome. In contrast, the PEB method overstates the policy gains in the various long-run outcomes by a much larger margin in the LAUSD, in the range 24-26 percent. The overestimation in both datasets is very similar to our short-run findings above that used mathematics test scores.

8 Extensions

Given the structure of our data, in which each teacher teaches one class of students each year – a feature common to many education datasets – our methodology can be extended to allow for *either* class-level shocks or drift in teacher quality: it does not allow us to account for both simultaneously.[50] The amount of drift in both North Carolina and LAUSD appears limited, consistent with prior research.[51] We focus on extending our model to allow for class-level shocks rather than drift, as these are quantitatively more important in our data.[52]

Class-level Shocks: These are shocks affecting everyone in a given classroom. The proverbial example involves a dog barking outside a given classroom window on test day, lowering the test scores of all students in that class; because teachers are unable to control the dog barking, the class-level shocks should not be attributed to the teacher. Accounting for such shocks reduces dispersion in teacher VA distribution, as they subsume some of the class-year variation that was previously attributed to teachers. Since the variance of the VA distribution 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 such a pattern.

Given that classrooms in our data are identified by unique teacher-year pairs, we rewrite our model given by equation (2.2) as:

$$y_{ijt} = \alpha_j + \theta_{jt} + \epsilon_{ijt}, \quad i = 1, 2, \ldots, n_{jt}, \tag{8.1}$$

[49]The explanation is as follows: given the truncation in the left tail, our non-parametric method uncovers a less dispersed distribution and so finds a weaker relationship between a one standard deviation increase in VA and long-run outcomes.

[50]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.

[51]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.

[52]While accounting for drift is beyond the scope of this paper, we note that is it likely feasible to do so with some recent notable advances being made in this direction – see Gourieroux and Jasiak (2020) for instance.

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^2_θ, it follows that the teacher-year specific sample mean can be modeled as $y_{jt} = \alpha_j + \nu_{jt}$ where $\nu_{jt} \sim N(0, \sigma^2_\theta + \frac{\sigma^2_\epsilon}{n_{jt}})$. Once again, the assumption that classroom and student-level shocks are normally distributed is not necessary (see Appendix B.2 for proof), although it is imposed for estimation purposes. The teacher-specific fixed effect estimator y_j is then constructed as before (see equation (2.3)), except the weights h_{jt} now become $\left(\sigma^2_\theta + \frac{\sigma^2_\epsilon}{n_{jt}}\right)^{-1}$. The NPB estimator, parallel to the result in Theorem 2, can be expressed as follows:

$$\delta_j^{NPB} = y_j + \Big(\sum_t \Big(\sigma^2_\theta + \frac{\sigma^2_\epsilon}{n_{jt}}\Big)^{-1} \Big)^{-1} \frac{\partial}{\partial y} \log g_j(y)|_{y=y_j} \,, \tag{8.2}$$

where $g_j(\cdot)$ is the marginal density of y_j. (See Appendix B.3 for formal derivation.) As in the case without classroom shocks, the feasible version of this estimator (i.e., the NPEB estimator) can be constructed by estimating F using NPMLE and the parameters σ^2_θ and σ^2_ϵ, via maximum likelihood (as described in Appendix F.2).

Results: Having incorporated classroom shocks in our estimation of the teacher quality distribution, adding them reduces the variance of our estimated distribution, as expected, with this reduction being more pronounced in the LAUSD data relative to the North Carolina data. In terms of policy, the reduction in variance acts to reduce the gains from teacher release policies since the quality of left tail teachers improves. The increased quality of left tail teachers is also likely to reduce the overestimation of PEB in the LAUSD, since the extra mass of left tail teachers in the PEB relative to NPEB will be of higher quality in a distribution that features lower variance.

Revisiting the policy evaluation, Figure I.2 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. Relative to the model without classroom shocks, overestimation of the PEB in both datasets is reduced since the addition of classroom shocks reduces the variance in the PEB methodology by more than in the NPEB methodology; further, 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.

Under the benchmark policy that release the bottom 5% of teachers, we find that the test score gains among affected students in North Carolina are 0.31 under both PEB and NPEB (in comparison, the PEB slightly overestimated gains by five percent in the model without class-level shocks). In contrast, test score gains in the LAUSD are 0.40 and 0.46 under the NPEB and PEB, respectively.[53] The PEB thus significantly *overestimates* policy gains in the LAUSD, by sixteen percent, relative to the twenty-six percent overestimation of PEB in the model without class-level shocks. The inclusion of class-level shocks therefore has little impact on the policy differences

[53]Exact numbers along with standard errors are: 0.399 (s.e. 0.007) for the NPEB and 0.462 (s.e. 0.006) for the PEB.

betwen the PEB and NPEB methodologies: Under our benchmark policy that releases bottom-5 percent teachers, the PEB and NPEB method estimate similar policy gains in North Carolina, while the PEB continues to overestimate policy gains in the LAUSD, by a significant amount.

9 Conclusion

In this paper, we have proposed a nonparametric approach for estimating teacher VA, relaxing the normality assumption embedded in the popular parametric empirical Bayes method. Our nonparametric empirical Bayes estimator is appealing in that it allows the underlying distribution of teacher quality to be estimated directly and in a computationally feasible way.

We documented that our estimator performs very well in simulations, being responsive to the underlying value-added distribution in ways the parametric method is not. The nonparametric estimator also has appealing finite sample properties, as shown in a variety of simulations. We applied the methodology using two separate administrative datasets in education, finding that the estimated teacher VA distributions differed from each other and departed from normality. We then explored the implications of these departures in a range of policy evaluations, showing that the benefits of teacher lay-off policies may be overstated to a large degree (in one of the two settings).

The nonparametric approach to estimation has broader applicability in 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. Suppose that policymakers implemented a policy releasing teachers at 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 repeated applications of the policy. Given that our method estimates changes like these in the underlying teacher quality distributions flexibly, it should provide more accurate predictions regarding the 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 datasets become more widely available in various fields, so the range of feasible applications using the nonparametric empirical Bayes approach is also likely to increase. To that end, we have written, and are making available, code that will allow researchers to implement the NPEB method in applications such as

these.

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.

Gourieroux, Christian and Joann Jasiak (2020), "Dynamic deconvolution of (sub)independent autoregressive sources." URL `http://www.jjstats.com/papers/dynamdec.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.

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 (2020), "On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means." *Electronic Journal of Statistics*, 14, 2272–2297.

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.

Lindsay, Bruce G. (1995), *Mixture Models: Theory, Geometry, and Applications*. Conference Board of the Mathematical Sciences: NSF-CBMS regional conference series in probability and statistics, Institute of Mathematical Statistics.

Los Angeles Unified School District (2003-2017), "Student, teacher, and demographic files." URL `https://achieve.lausd.net/research`.

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.

North Carolina Education Research Data Center (1996-2017), "Student, class and personnel files." URL `http://childandfamilypolicy.duke.edu/research/hc-education-data-center/`.

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.

Polyanskiy, Yury and Yihong Wu (2020), "Self-regularizing property of nonparametric maximum likelihood estimator in mixture models." *arXiv preprint arXiv:2008.08244*.

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 (1950), "A generalization of the method of maximum likelihood: Estimating a mixing distribution." *Annals of Mathematical Statistics*, 21, 314–315.

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.

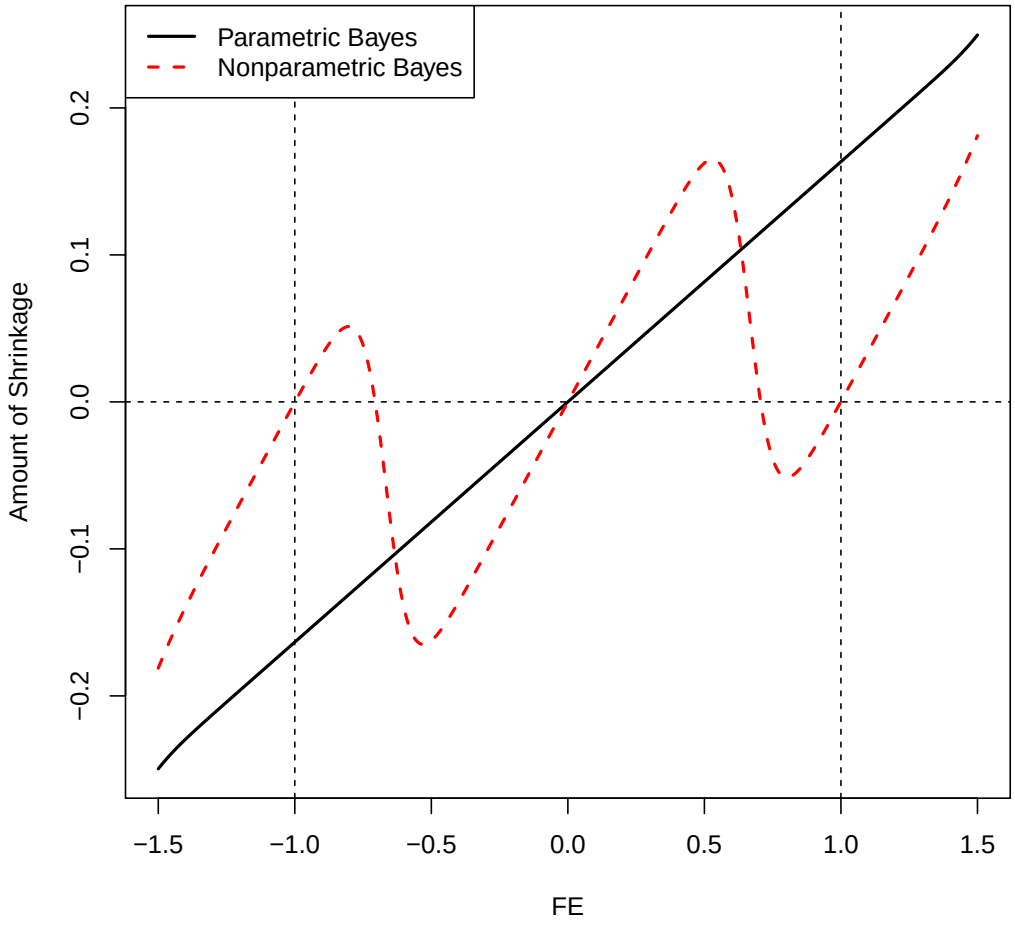
Saha, Sujayam and Adityanand Guntuboyina (2020), "On the nonparametric maximum likelihood

estimator for gaussian location mixture densities with application to gaussian denoising." *Annals of Statistics*, 48, 738–762.

van de Geer, Sara (1993), "Hellinger-consistency of certain nonparametric maximum likelihood estimators." *Annals of Statistics*, 14–44.

Zhang, Cun-Hui (2009), "Generalized maximum likelihood estimation of normal mixture densities." *Statistica Sinica*, 1297–1318.

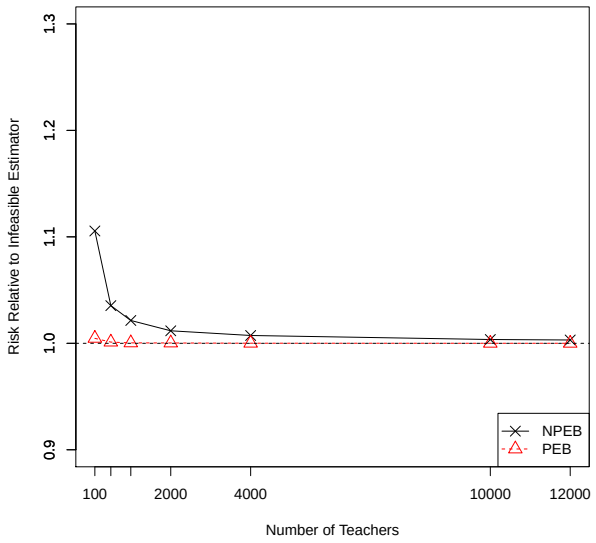
Figure 1: Example of Shrinkage under Parametric and Nonparametric Bayes Estimators



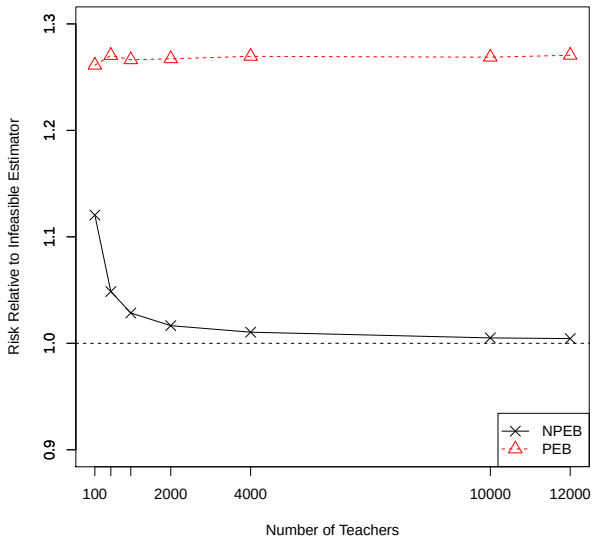
Notes: This figure plots the amount of Bayesian 'shrinkage' as a function of the fixed effect estimates (on the x-axis) for the parametric Bayes (PB) and nonparametric Bayes (NPB) estimators, respectively. The shrinkage rule for the PB estimator is given by $\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2/\sum_t n_{jt}}$ from equation (2.5) with $\sigma_\alpha^2 = 0.08$. For the NPB estimator, the shrinkage rule is given by the second term in equation (2.7). We compare shrinkage under the PB and NPB estimators assuming an underlying distribution for the teacher fixed effects α_j given by the mixed normal distribution $\alpha_j \sim 0.95\mathcal{N}(0, 0.03) + 0.025\mathcal{N}(-1, 0.03) + 0.025\mathcal{N}(1, 0.03)$. The total class size for each teacher is set at sixteen and σ_ϵ^2 is set at 0.25. Fixed effects take values in the range $[-1.5, 1.5]$. 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: Simulation Performance in Finite Samples (using MSE) for PEB and NPEB Estimators Relative to the Infeasible Benchmark when Teacher Quality Follows:

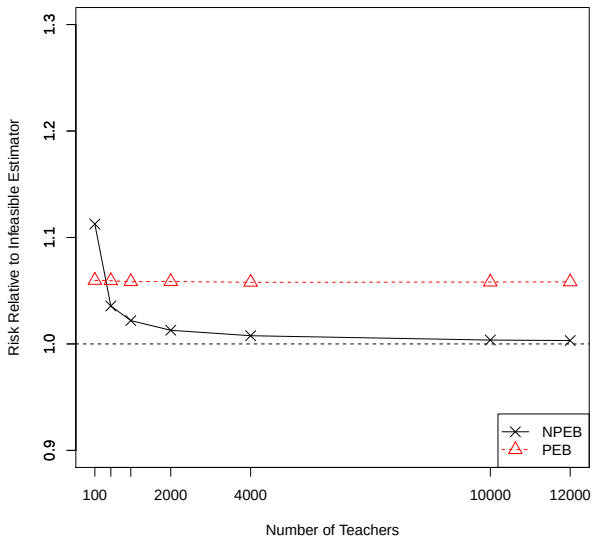
(a) Normal Distribution



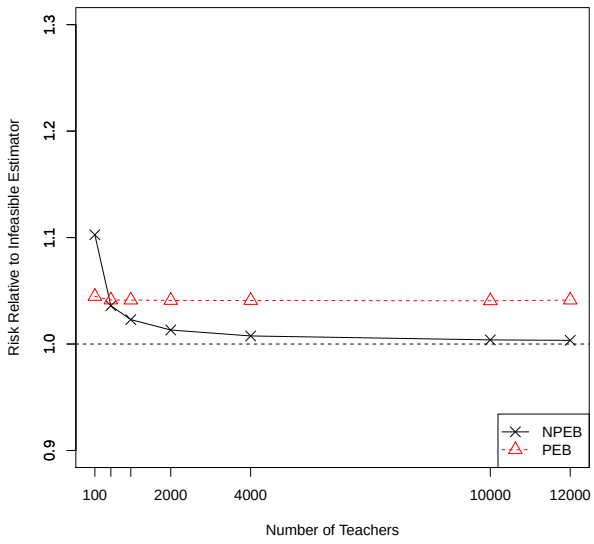
(b) Mixed Normal Distribution



(c) Skewed Normal Distribution

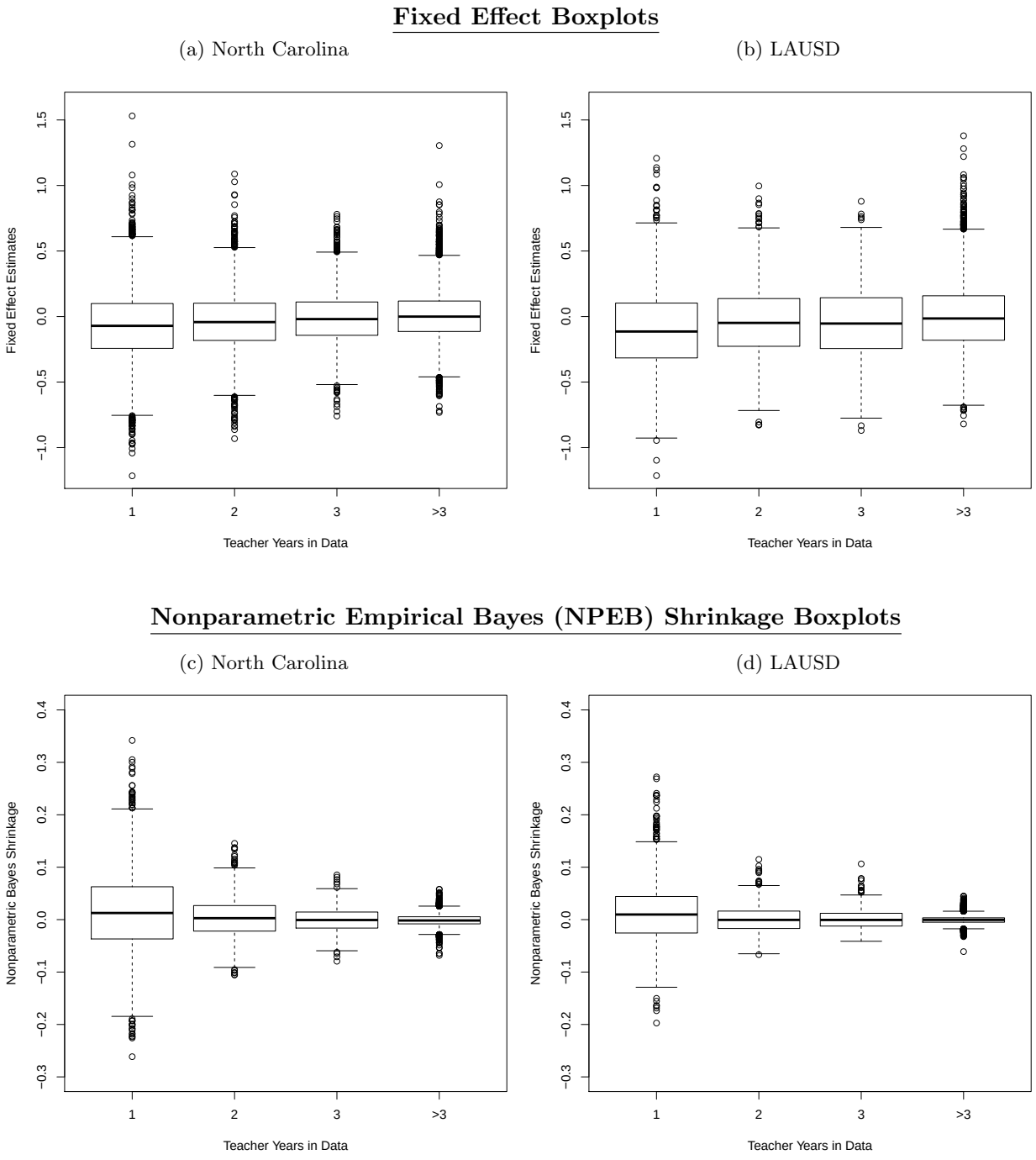


(d) Skewed Logistic Distribution



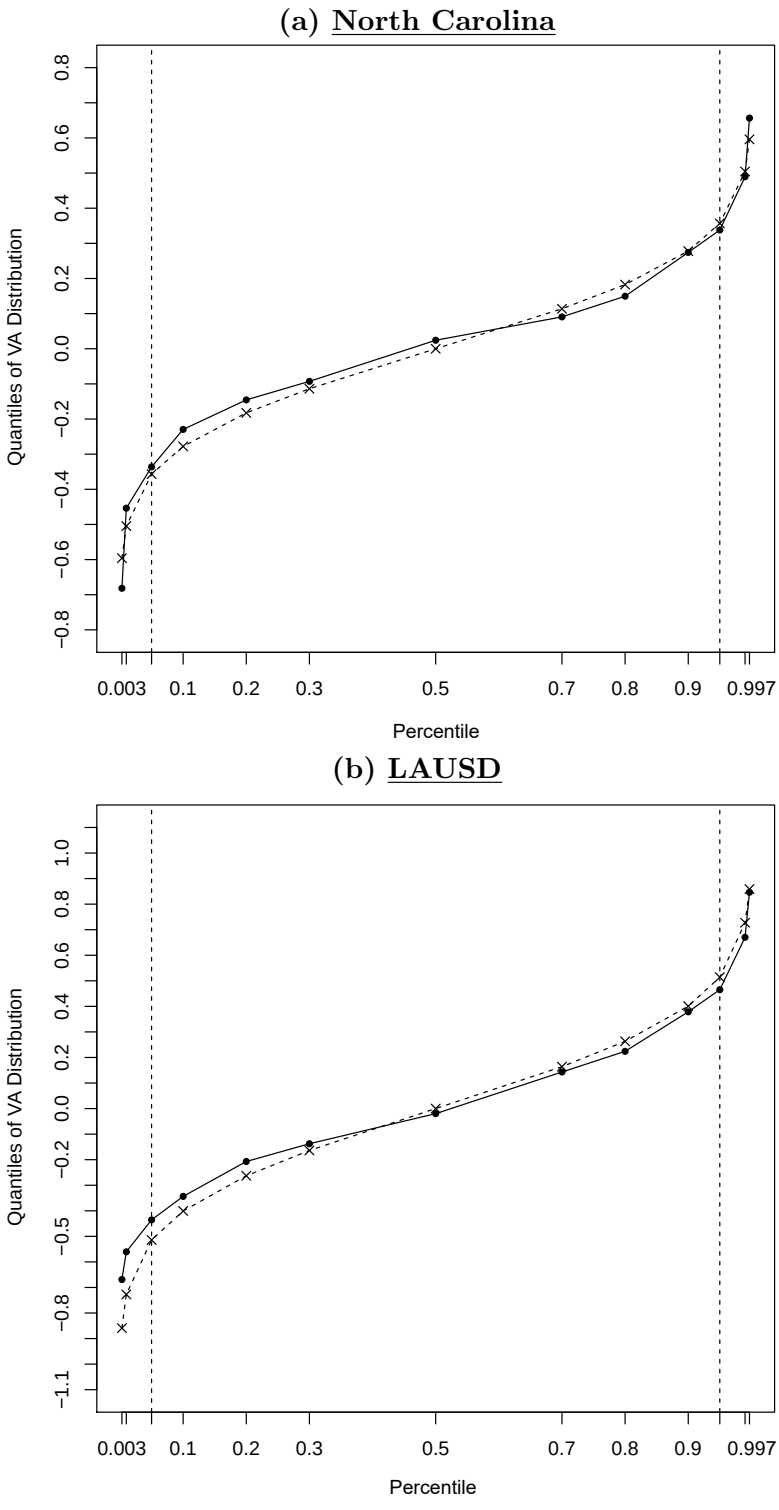
Notes: These figures assess the performance of the parametric (PEB) and our nonparametric (NPEB) estimator in finite samples relative to the infeasible benchmark for four data generating processes considered in our simulations. Each figure gives the mean squared error of the PEB and NPEB estimators reported as a risk ratio, which benchmarks their mean squared error relative to the infeasible estimator: a value of one implies the estimator has the same mean squared error as the infeasible estimator. Figure 2(a) simulates the risk ratio when the underlying distribution is normally distributed $F \sim \mathcal{N}(0, 0.08)$, and Figure 2(b) does so when the underlying distribution is mixed normal $F \sim 0.95\mathcal{N}(0, 0.03) + 0.025\mathcal{N}(-1, 0.03) + 0.025\mathcal{N}(1, 0.03)$. Figure 2(c) then skews the normal distribution to the right with location parameter -0.4 and shape parameter 5, while Figure 2(d) uses a skewed logistic with location parameter -0.5 and shape parameter 5. The parameters are chosen such that all four distributions are mean zero and have roughly the same variance. The x-axis reports number of teachers, each with a class size drawn from the set {8,16}. The simulations average results from 500 repetitions, setting $\sigma_\epsilon^2 = 0.25$.

Figure 3: Boxplots of Fixed Effects and of Shrinkage under Nonparametric Empirical Bayes



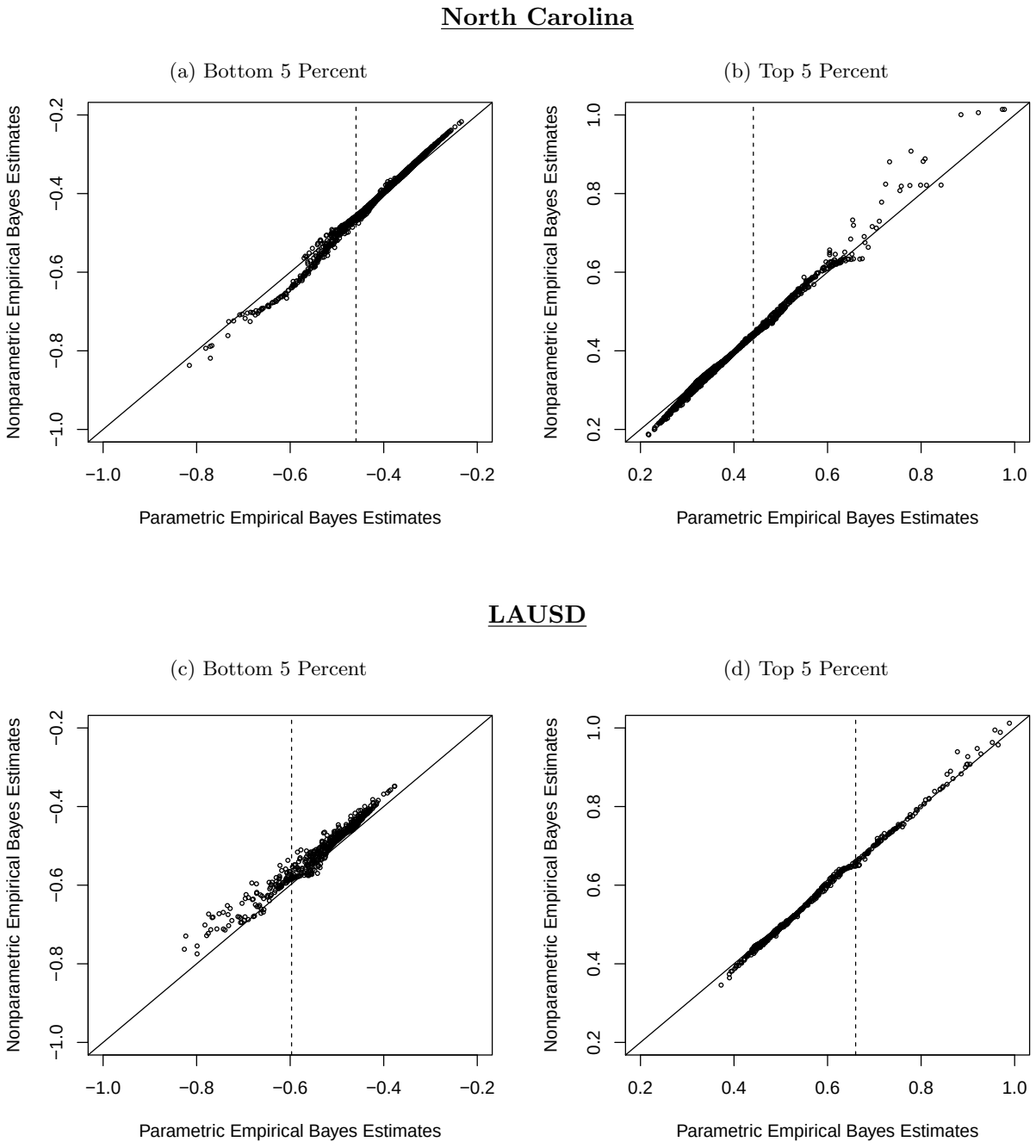
Notes: Figures 3(a) and 3(b) give the raw fixed effect estimates by the number of times a teacher appears in our North Carolina and LAUSD datasets, respectively. Specifically, each panel displays a boxplot of fixed effect estimates for teachers who appear once, twice, three times, and more than three times. (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.) Figures 3(c) and 3(d) then show boxplots of the amount of shrinkage applied by our NPEB estimator to teachers who appear once, twice, three times, and more than three times in our North Carolina and LAUSD datasets, respectively.

Figure 4: Estimated Teacher VA Quantile Functions using NPMLE Compared with Estimated Normal



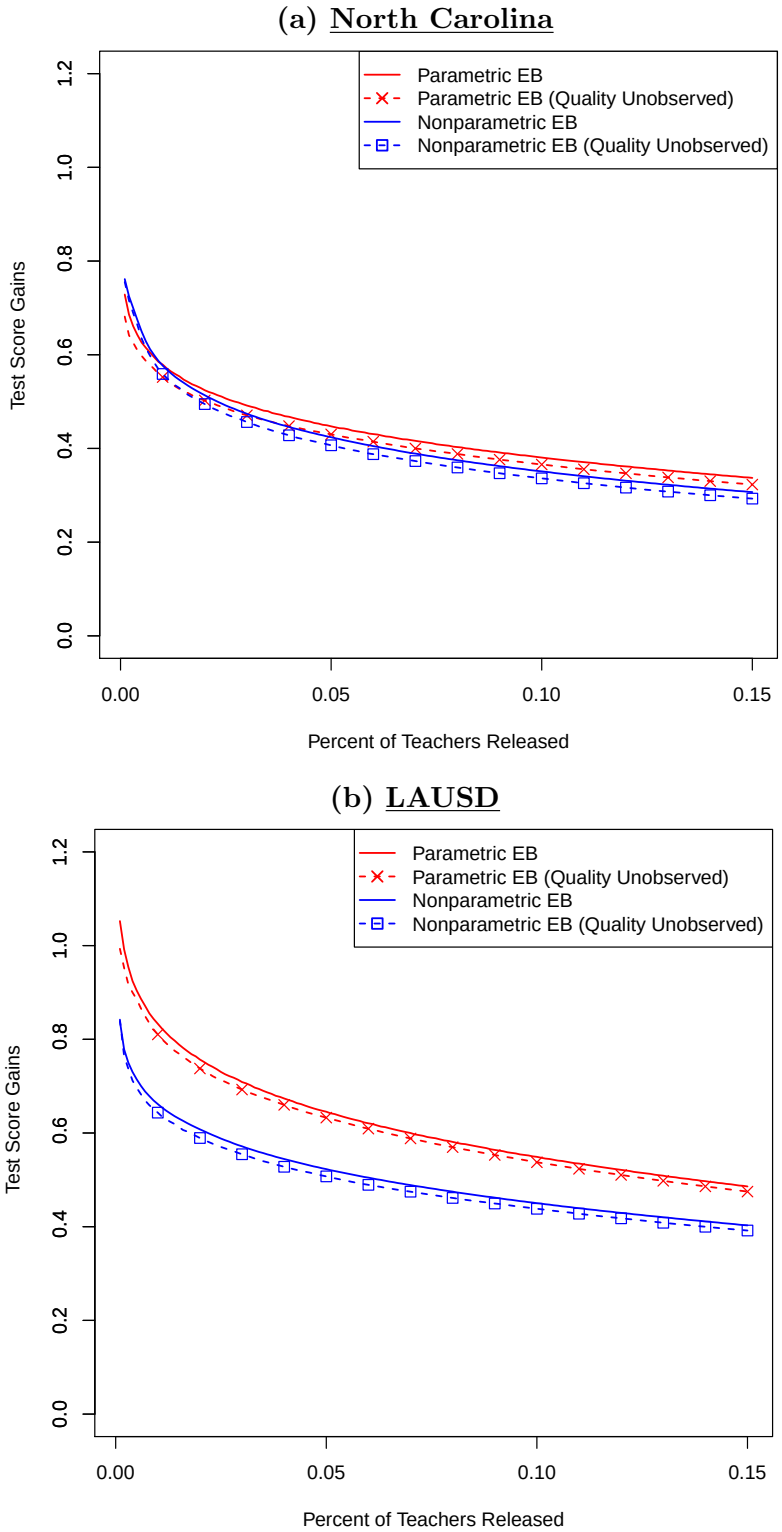
Notes: The panels in this figure show estimated quantile functions of teacher quality (VA) using the North Carolina and LAUSD datasets, respectively. Underlying values are provided in Table I.3. The solid line in each panel connects the estimated quantile values of the teacher quality distribution estimated via nonparametric maximum likelihood (see equation (3.1)) at different points in the distribution. The dashed line connects the estimated quantile values when normality is imposed (as in parametric empirical Bayes) estimated via maximum likelihood. (See Appendix F for more details.) The vertical dashed lines in each figure indicate the bottom and top 5 percentiles of the teacher quality distribution.

Figure 5: Parametric and Nonparametric Empirical Bayes 'Shrinkage' at the Bottom and Top Five Percent of the Fixed Effect Distribution



Notes: The panels in this figure show VA estimates calculated using the parametric and nonparametric empirical Bayes methodologies (the former on the x-axis, the latter on the y-axis) using our two datasets. Each dot represents a teacher. The diagonal line, the 45-degree line, indicates points where the two empirical Bayes estimates agree. The two panels to the left contrast the VA estimates among teachers whose estimated fixed effects are in the bottom five percent, while those on the right side do so for teachers in the top five percent. The vertical dashed lines in the two panels to the left mark the boundary showing the bottom one percent of the observations (based on PEB); the vertical dashed lines in the two panels to the right mark the boundary showing the top one percent (on the same basis).

Figure 6: Test Score Gains from Replacing the Bottom q Percent of Teachers



Notes: Figures 6(a) and 6(b) show the average test score gains among affected students of a policy that releases the bottom q% of teachers (replacing with mean-quality teachers) in North Carolina and the LAUSD, respectively. The dashed lines represent the policy gains expected under the PEB and NPEB methodology when true teacher VA is estimated, while the solid lines indicate the policy gains if true teacher VA were observed. Policy gains when teacher VA is estimated 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.) The numbers underlying the figures are the same as those reported in Table 4 when VA is estimated, and Table I.4 when true VA is observed.

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

	Homogeneous Class Size (Class Size of 16)				Heterogeneous Class Sizes (Class Size 8-16)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
SSE	130.50	131.07	130.51	156.06	186.30	186.96	186.31	233.99
Bias in Bottom 5%	0.000	0.000	0.000	-0.104	0.000	0.000	0.000	-0.140

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

	Homogeneous Class Size (Class Size of 16)				Heterogeneous Class Sizes (Class Size 8-16)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
SSE	106.8	107.5	130.5	156.1	140.0	140.6	177.5	234.0
Bias in Bottom 5%	0.000	0.000	0.037	-0.085	0.000	0.000	0.051	-0.125

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

	Homogeneous Class Size (Class Size of 16)				Heterogeneous Class Sizes (Class Size 8-16)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
SSE	128.5	129.1	135.2	156.0	176.7	177.3	187.0	234.0
Bias in Bottom 5%	0.000	0.000	-0.075	-0.150	0.000	0.000	-0.089	-0.207

Table 1(d): Simulation – *True Distribution is Skewed Logistic*

	Homogeneous Class Size (Class Size of 16)				Heterogeneous Class Sizes (Class Size 8-16)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
SSE	132.2	132.9	136.4	156.3	181.8	182.5	189.2	234.3
Bias in Bottom 5%	0.000	0.000	-0.047	-0.124	0.000	0.000	-0.059	-0.178

Notes: The panels in this table report simulation results comparing the performance of the three candidate estimators (across the columns) against an infeasible benchmark, based on the sum of squared errors ('SSE') and selection bias in the bottom 5 percent of the distribution. The infeasible benchmark is the optimal empirical Bayes 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. The precise specifications of the normal distribution in Table 1(a), the mixed normal in Table 1(b), the skewed normal in Table 1(c), and the skewed logistic in Table 1(d) are described in Section 4. The simulations average results from 500 repetitions with 10,000 individual teachers, setting $\sigma_\epsilon^2 = 0.25$. Results are reported on the left side of each panel for homogeneous class sizes (where every teacher has a class size of sixteen) and on the right side, for heterogeneous class sizes (where class sizes are drawn randomly from the set $\{8, 16\}$ with equal probability). The formulae for risk ('SSE') and selection bias ('Bias in the Bottom 5%') are also given in Section 4. (Tables I.1(a)-I.1(d) expand on the tables here by also providing selection bias in the top 5% for these simulations, while Tables I.2(a)-I.2(d) give type I and II error rates for these simulations.)

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				
Mathematics Score (σ)	0.00	0.05	0.00	0.07
Reading Score (σ)	0.00	0.03	0.00	0.06
Lagged Mathematics 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 (student-year)[6]	4,457,812	2,680,027	1,707,459	1,280,569

Notes: this table presents summary statistics for the variables in our North Carolina and Los Angeles Unified School District (LAUSD) datasets, respectively, comparing the full and value-added samples.

[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 comparing columns (1) and (2) arises because we drop 1.37 million student-year observations that cannot be matched to their classroom teacher (see Appendix C for more detail).

[2] 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] This variable is missing for school years 1996-97 through 1997-98 in North Carolina.

[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 around 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

	NMSE			TMSE		
# of Prior Years Used	NPEB	PEB	FE	NPEB	PEB	FE
$t = 1$	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

	NMSE			TMSE		
# of Prior Years Used	NPEB	PEB	FE	NPEB	PEB	FE
$t = 1$	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: 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 sizes that teachers face across years, we use normalized mean squared error (NMSE) and total mean squared error (TMSE), as proposed by Brown (2008): see equation (6.1) and equation (6.2), respectively. The prediction performance is calculated by computing 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. Smaller values indicate better prediction performance.

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

% Teachers Released (q)	North Carolina Data		LAUSD Data	
	Score Gain under F (NPEB)	Score Gain under Normal (PEB)	Score Gain under F (NPEB)	Score Gain under Normal (PEB)
	(1)	(2)	(3)	(4)
1	0.559	0.558	0.643	0.814
	(0.012)	(0.006)	(0.014)	(0.010)
3	0.456	0.478	0.555	0.698
	(0.006)	(0.004)	(0.008)	(0.007)
5	**0.407**	**0.437**	**0.507**	**0.638**
	(0.005)	(0.003)	(0.007)	(0.006)
7	0.373	0.407	0.475	0.595
	(0.004)	(0.003)	(0.006)	(0.005)
9	0.347	0.382	0.450	0.559
	(0.004)	(0.003)	(0.005)	(0.005)

Notes: Table 4 displays the mathematics score gains of a policy that releases the bottom q% of teachers and replaces them with average-quality teachers. The gains are among affected students in terms of student-level standard deviations. True teacher quality is taken to be unobserved by the policymaker, the gains being the same as those in panels (a) and (b) of Figure 6 for 'Quality Unobserved.' 'Score Gain under F' reports the test score gain of the policy when teacher quality is distributed according the distribution F, estimated nonparametrically using equation (3.1), and applying the NPEB estimator to calculate VA. 'Score Gain under Normal' reports the test score gain when teacher quality is normally distributed and the PEB estimator is used to calculate teacher VA. The bolded line indicates the widely-analyzed 'release the bottom five percent of teachers' policy. 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. Standard errors are calculated using the bootstrap as described in Appendix E. (This table is comparable to Table I.4 which reports policy gains when true teacher quality is observed by the policymaker, and Figure I.2 which adds classroom shocks to the teacher VA model.)

Table 5(a): Long-Run Gains from Releasing Bottom 5% of Teachers – True VA Unobserved (North Carolina)

Long-Run Outcome:	Percent Drop Out (1)	Days Suspended (2)	PSAT Score (3)	Percent Took SAT (4)	SAT Score (5)	Exit Exam Score (6)
Panel A. Parametric Empirical Bayes						
Benefit ($\hat{\kappa}^{PEB}$)	-0.36	-0.065	6.63	1.01	6.33	-
Average Change in VA of Released Teachers ($\hat{\Delta}m_{\sigma}^{PEB}$)	1.98	1.98	1.98	1.98	1.98	-
Gain of Releasing Bottom 5% (G^{PEB})	-0.71	-0.130	13.13	1.99	12.53	-
Panel B. Nonparametric Empirical Bayes						
Benefit ($\hat{\kappa}^{NPEB}$)	-0.33	-0.061	6.18	0.94	5.84	-
Average Change in VA of Released Teachers ($\hat{\Delta}m_{\sigma}^{NPEB}$)	2.03	2.03	2.03	2.03	2.03	-
Gain of Releasing Bottom 5% (G^{NPEB})	-0.67	-0.124	12.54	1.90	11.86	-
Overestimation of PEB (%)	4.9	4.6	4.7	4.7	5.7	-

Table 5(b): Long-Run Gains from Releasing Bottom 5% of Teachers – True VA Unobserved (LAUSD)

Long-Run Outcome:	Percent Drop Out (1)	Days Suspended (2)	PSAT Score (3)	Percent Took SAT (4)	SAT Score (5)	Exit Exam Score (6)
Panel A. Parametric Empirical Bayes						
Benefit ($\hat{\kappa}^{PEB}$)	-0.28	-0.012	8.74	0.27	6.65	2.74
Average Change in VA of Released Teachers ($\hat{\Delta}m_{\sigma}^{PEB}$)	2.02	2.02	2.02	2.02	2.02	2.02
Gain of Releasing Bottom 5% (G^{PEB})	-0.57	-0.024	17.65	0.55	13.44	5.53
Panel B. Nonparametric Empirical Bayes						
Benefit ($\hat{\kappa}^{NPEB}$)	-0.24	-0.010	7.63	0.24	5.78	2.39
Average Change in VA of Released Teachers ($\hat{\Delta}m_{\sigma}^{NPEB}$)	1.87	1.87	1.87	1.87	1.87	1.87
Gain of Releasing Bottom 5% (G^{NPEB})	-0.45	-0.019	14.26	0.45	10.81	4.47
Overestimation of PEB (%)	26.3	24.8	23.8	24.2	24.3	23.9

Notes: Tables 5(a) and 5(b) show – using the North Carolina and LAUSD data, respectively – the estimated gains in terms of various long-run outcomes of a policy that releases the bottom 5% of teachers and replaces them with mean quality teachers when true teacher quality is *unobserved* by the policymaker. The 'Benefit' row in each panel represents the increase in the long-run outcome associated with having a teacher whose VA is one standard deviation higher. These rows (along with the first row of sample means) are identical to the corresponding rows in Tables I.5(a) and I.5(b). Differences appear in the third row in each panel, however, which calculates the average change in VA caused by releasing bottom 5% teachers since some teachers are mistakenly released as true VA is not observed as in Tables I.5(a) and I.5(b). Here, we calculate the average VA of the bottom 5% teachers 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 in each year. This average VA of the bottom 5% teachers is then multiplied by the benefit to give the policy effect of releasing bottom 5% teachers according to *estimated* VA (given by third row of each panel). The final row of each table, 'Overestimation of Parametric EB,' gives the overestimation of policy gains in terms of the long-run outcome from utilizing PEB rather than NPEB. Long-run outcomes are: high school drop out, total days suspended in middle and high school, PSAT scores, whether student takes the SAT, and SAT scores, and exit exam scores (only available for LAUSD). For PSAT and SAT scores, we combine the mathematics and English components and take the values from the student's first attempt. (See Appendix C.3 for more details about data construction.) Tables I.5(a) and I.5(b) repeat an analogous exercise when true teacher quality is observed to the policymaker and so teacher releases are based on true (rather than estimated) value-added: the results from doing so are virtually identical.

A Proof of Theorems

Proof. [Theorem 1] 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. ■

Proof. [Theorem 2] 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, the nonparametric Bayes estimator is given by

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

■

B 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.

B.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 function $\phi_\alpha(\cdot)$ of α and the characteristic function $\phi_\epsilon(\cdot)$ for ϵ_1 and ϵ_2 are nonvanishing everywhere.*

Lemma 1 (Kotlarski (1967)) *Under Assumptions 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 $\phi_\epsilon(t)$ be 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

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

We then apply the inverse Fourier transform to $\phi_\alpha(t)$ and $\phi_\epsilon(t)$, yielding the density functions of α and ϵ.

Corollary 3 *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_{js} (with $j = 1, 2, \ldots, J$, $s = 1, 2, \ldots, n_j$) are random variables from the same pdf $h(\cdot)$ with mean zero. If $n_j \geq 2$, α and ϵ_{js} are mutually independent, and the characteristic functions $\phi_\alpha(\cdot)$ for α and $\phi_\epsilon(\cdot)$ for ϵ_{js} are nonvanishing everywhere, then the pdf's of α and ϵ are nonparametrically identified.

The above corollary applies to 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 then construct the nonparametric estimator for $\psi(\cdot,\cdot)$ as

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

B.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_1, θ_2 and ϵ_{ij} for $\{i,j\} \in \{1,2\}$ are nonvanishing everywhere.*

Lemma 2 *Under Assumptions 4-6, the pdf's of α, θ_1, θ_2 and ϵ_{ij} for $\{i,j\} \in \{1,2\}$ 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 to

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

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.3 Deriving NPEB with Classroom Shocks

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 4 *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 for 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, this is not necessary in Appendix B.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 F.2).

C 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 datasets. 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.

C.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 the state in grades 4-5 for the 1996-97 through 2010-11 school years and grade 3 for the 1996-97 through 2009-10 school years.[54] These data cover roughly 1.85 million students with 4.5 million student-year observations.

In terms of 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 missing data in some years for parental education and economically disadvantaged status, our demographic data cover over 99 percent of all student-year observations. When 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 suspect as having an invalid teacher; this is by far our biggest sample restriction.[55] Second, charter school classrooms and special education classrooms are dropped, leading to a loss of an additional 70,000 student-year observations. Third, we drop 16,000 observations where teacher experience data are missing. Fourth, 380,000 observations lacking a valid current or lagged test score in that subject are excluded, 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.[56] Fifth, we only include classes with more than seven but fewer than forty students with valid current and lagged test scores in that subject, resulting in a loss of 10,000 observations.[57] Our final sample consists of roughly 2.7 million student-year observations, covering 1.4 million students and 35,000 teachers.

[54] 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.

[55] We assign teachers to students based on the person recorded as proctoring the student's exam, further confirming that the proctor is teaching a primary grade mathematics and English class. If the teacher is not, we drop the observations.

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

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

C.2 Los Angeles Unified School District (LAUSD)

For the LAUSD dataset, we start with the entire enrollment history of students in the district in grades 4-5 for 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 since we only have teacher experience data from 2007-08 onwards.[58] 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.[59] 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, 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

[58] 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.

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

and school-year means of all the individual covariates, X_{igt}, and (iv) grade and year dummies.

C.3 Long-Run Outcome Data

We describe the long-run outcome data we merge onto our main analysis dataset briefly. For both the LAUSD and North Carolina data, we focus on five outcomes: high school dropout status, days suspended, PSAT scores, SAT taking, and SAT scores. In addition, we use scores on the state standardized exit exam (CAHSEE) as an additional outcome in the LAUSD data. We explain how we merge these long-run outcomes into each dataset below.

LAUSD: California High School Exit Examination (CAHSEE) data cover 2003-04 through 2014-15. As the CAHSEE is normally first taken in tenth grade, we keep cohorts who were in tenth grade during this time period.[60] If a student took the CAHSEE multiple times, we take the score from the student's first sitting of the CAHSEE. We report the sum of CAHSEE scores from the mathematics and English sections and so CAHSEE scores range from 550 to 900. We have CAHSEE records for 184,128 students, covering sixty-seven percent of students from eligible cohorts.

Data on dropouts and graduations are available for school years 2011-12 through 2016-17. These data indicate whether students in the twelfth grade cohort of that year graduated or dropped out and so we restrict our data to cohorts in twelfth grade during this time period.[61] We have a dropout or graduation record for 129,456 students – fifty percent of students from eligible cohorts.

Suspension data are available from 2003-04 through 2016-17. We restrict data to 2003-04 through 2009-10 to allow for sufficient time to elapse for students to receive suspensions after a student is taught by a given teacher. We calculate the total number of days of out-of-school suspensions for each student occurring in middle or high school (grades 6-12). If we do not find the student in the suspension files, we assume the student has never been suspended. Of note, the LAUSD embarked on an ambitious policy to eliminate "wilful defiance" suspensions, causing a large drop (almost seventy-five percent) in suspension rates in the 2010s, creating a low rate of suspension in the LAUSD data. We have suspension outcomes for our full value-added sample of 426,074 students from 2003-04 through 2009-10.

PSAT data cover school years 2008-09 through 2016-17. As the PSAT is normally taken in tenth grade, we keep cohorts who were in tenth grade during this time period.[62] We convert PSAT scores from the 2015-16 and 2016-17 administrations using the concordance tables provided by the College Board so that all PSAT scores are reported on a 600-2400 scale.[63] If a student is recorded as receiving multiple administrations of the PSAT, we take the score from the first PSAT test taken by the student. We have PSAT records for 209,675 students, covering fifty-one percent of students from eligible cohorts.

SAT data cover school years 2006-07 through 2016-17. As the SAT is usually taken in the eleventh or twelfth grade, we keep cohorts who attended both grades during the time period covered

[60] We therefore drop fifth grade after 2009-10, fourth grade after 2008-09, and third grade after 2007-08.
[61] We therefore drop fifth grade after 2008-09, fourth grade after 2007-08, and third grade after 2006-07.
[62] We therefore drop fifth grade after 2011-12, fourth grade after 2010-11, and third grade after 2009-10.
[63] Available at `https://collegereadiness.collegeboard.org/pdf/2015-psat-nmsqt-concordance-tables.pdf`.

by our SAT data.[64] SAT scores are recorded on a 400-1600 scale. As with the PSAT, we only keep the student's first score if they have taken the test multiple times. We have SAT records for 92,309 students – thirty percent of students from eligible cohorts.

North Carolina: Dropout data are available for school years 2003-04 through 2016-17. Given the majority of students drop out in tenth through twelfth grade, we ensure that our dropout data coverage starts in at least tenth grade for a given cohort[65] and covers up to twelfth grade.[66] Any student that has either not dropped out or moved out-of-state is coded as not being a high school dropout. We have dropout outcomes for 1,097,381 students.

Suspension data are available from 2000-01 through 2016-17, although data from 2004-05 are missing. Here, we keep all data, since the suspension data cover a majority of high school years for all cohorts. We find the total number of days of out-of-school suspensions for each student occurring in middle or high school (grades 6-12). We top-code the number of days suspended in a school year at ten days. If we do not find the student in the suspension files, we assume that the student has never been suspended. Thus we have suspension outcomes for our full value-added sample of 1,386,555 students.

PSAT data cover school years 2012-13 through 2016-17. As the PSAT is normally taken in tenth grade, we keep cohorts in tenth grade from 2012-13 through 2016-17.[67] We convert PSAT scores from the 2015-16 and 2016-17 administrations using the concordance tables provided by the College Board, so that all PSAT scores are reported on a 600-2400 scale. If a student is recorded as receiving multiple administrations of the PSAT, we take the score from the first time the student took the PSAT test. We have PSAT records for 159,028 students, covering forty-four percent of students from eligible cohorts.

SAT data cover school years 2008-09 through 2016-17. As the SAT is usually taken in the eleventh or twelfth grade, we keep cohorts who attended both grades during the time period covered by our SAT data.[68] SAT scores are recorded on a 400-1600 scale. As with the PSAT, we only keep the student's first score if they have taken the SAT multiple times. We have SAT records for 275,3810 students – thirty-five percent of students from eligible cohorts.

[64] We therefore drop fifth grade after 2009-10, fourth grade after 2008-09, and third grade after 2007-08.

[65] This means we drop fifth grade students in 1996-97 and 1997-98 and fourth grade students in 1996-97.

[66] This necessitates that we drop 2010-11, fourth grade students in 2009-10, and third grade students in 2008-09.

[67] We thus keep fifth grade from 2007-08, fourth grade from 2006-07, and third grade from 2005-06. Fourth grade in 2010-11 is also dropped (recall third grade coverage ends in 2008-09 in the VA sample).

[68] The data are therefore restricted to third grade in 1999-00 through 2007-08, fourth grade in 2000-01 through 2008-09, and fifth grade from 2001-02 through 2009-10.

D Linking Long-run Outcomes to Teacher VA

In Section 7.2, we referenced a method to link long-run outcomes with teacher VA proposed by Chetty et al. (2014b). In this appendix, for completeness, we describe the steps involved.

The first step involves constructing long-run outcome residuals using variation across students taught by the same teacher j, based on the regression equation

$$Y_{ij}^{*} = \alpha_j + \beta^Y X_{ijt} + u_{ijt}\,, \tag{D.1}$$

where Y_{ij}^{*} is the long-run outcome of interest, α_j is a teacher fixed effect , and X_{ijt} are observed characteristics of the student and the teacher. Using the estimates from equation (D.1), the long-run residuals, Y_{ijt}, are defined as:

$$Y_{ijt} = Y_{ij}^{*} - \hat{\beta}^Y X_{ijt}\,. \tag{D.2}$$

We then regress these long-run residuals on each teacher's (normalized) VA, pooling across all grades, using VA estimates based on both the parametric and our nonparametric approaches:

$$Y_{ijt} = \delta + \kappa^d \hat{m}_{jt}^d + \eta_{ijt}\,, \;\; d \in \{PEB, NPEB\}, \tag{D.3}$$

where $\hat{m}_{jt}^d \equiv \hat{\alpha}_j^d / \hat{\sigma}_\alpha^d$ denotes 'normalized' teacher VA, which is our estimate of a teacher's VA (with the superscript denoting whether it is calculated using PEB or NPEB), scaled by the estimated standard deviation ($\hat{\sigma}_\alpha^d$) of the teacher VA distribution. (As in Chetty et al. (2014b), the standard deviation of the normalized VA measure is less than one since Bayes shrinkage is applied to the VA estimates.)

The impact of being assigned a teacher with higher VA for one year on long-run outcomes in North Carolina and the LAUSD are shown in appendix Figures I.3 and I.4, respectively. The panels plot residual long-run outcomes for students in school year t versus the estimated (normalized) VA using NPEB of the teacher j who taught them in that year, given by $\hat{m}_{jt}^{NPEB}$.[69] In terms of the slope coefficients underlying each panel, we find that being assigned to a teacher whose test score VA based on NPEB is one SD higher in a single grade in the LAUSD decreases the drop-out rate by 0.24 percentage points, lowers suspensions by 0.01 days, increases PSAT scores by 7.6 (on the 2400 scale), boosts SAT taking by 0.24 percentage points, increases SAT scores by 5.8 (on the 1600 scale), and raises high school exit exam scores by 3.6% relative to the mean. We find similar relationships in the North Carolina data. All of these increases are statistically significant at the one percent level.

[69] To construct the binned scatter plot, we take three steps: (i) residualize the long-run outcome as described in equation (D.1), (ii) divide our VA estimates, $\hat{m}_{jt}$, into twenty equal-sized bins and plot the mean of the long-run outcome residuals in each bin against the corresponding bin mean of $\hat{m}_{jt}$, and (iii) add back in the mean long-run outcome in the estimation sample to facilitate the interpretation of the scale.

E Bootstrapping the Standard Errors

This appendix describes how we generate the bootstrapped standard errors for the policy evaluations. Uncertainty in the policy analysis regarding test score gains under teacher release or retention policies stems from the fact that the distribution of the teacher quality – either nonparametrically identified from the data or under the parametric assumption of Gaussian – has to be estimated from 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: (i) 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). (ii) 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 $B = 800$ times, and calculate the standard errors 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 constructed from the empirical quantiles of the estimates of teacher VA, for each bootstrap distribution $\hat{F}^{(b)}$ conduct step (iii): 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 teachers 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: (i) 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). (ii) 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 (iii): 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.

F Maximum Likelihood Estimation of Variance Parameters

F.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 when 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 a method-of-moments estimator proposed by Kane and Staiger (2008) under the additional assumption that $\alpha_j \sim N(0, \sigma_\alpha^2)$. Specifically, they propose the following estimators for the variance parameters:

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

This is also the estimator used by Chetty et al. (2014a) for teacher VA without drift. The main shortcoming of the method-of-moments 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, and denoting 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-moments estimator, all individuals, including those with only one period of data, are accounted for here.

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

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

F.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 also 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-moments 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-moments 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. It can be shown 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 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 $y_{jt} \sim N(0, \sigma_\alpha^2 + \sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n_{jt}})$, and 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.

G 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. 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, based on 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, 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. The table 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 North Carolina and LAUSD data. For the North Carolina 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 are also possible. For instance, if α indeed follows a normal distribution $N(0, \sigma_\alpha^2)$, then the logarithm 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 has been proposed in Bonhomme and Weidner (2019). We leave to future research a power comparison involving these and other specification tests.

H Computation Appendix

In this brief appendix, we discuss the estimation of F using NPMLE.

H.1 Computing the NPMLE for F

The difficulty in estimating equation (3.1), as pointed out in Koenker and Mizera (2014), is that F is an infinite-dimensional object, involving an infinite number of constraints. To make computation feasible and maintain the convexity of the problem, those authors 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. 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.[70]

H.2 Constructing the NPEB Estimator

We construct a feasible version of the NPB estimator based on (2.6) directly, rather than its equivalent reformulation (2.7) in Theorem 2. We do so for two reasons. First, equation (2.7) suggests that the nonparametric Bayes estimator δ_j^{NPB} depends on the marginal density of the fixed effect estimator, rather than the teacher quality distribution F directly. Therefore, in principle, we could focus on constructing a feasible estimator for the marginal density, yet in practice, this becomes challenging when individual teachers have heterogeneous variances.[71]

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.[72] 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.

The proposed estimator (3.2) can be motivated on the grounds that it makes effective use of information from the data (to learn about F) and from the model (using the normal mixture structure). Saha and Guntuboyina (2020) recently showed that the NPEB estimator δ_j^{NPEB} constructed

[70] For the datasets we apply this method to, the sample size is around 35,000 for the North Carolina data and 11,000 for the LAUSD data. In both cases, we take $M = 5000$. Making M even larger in both examples does not further improve the precision.

[71] Brown and Greenshtein (2009) propose a kernel method to estimate the marginal density of y_j directly when variances of y_j are *all the same* – for example, when all teachers have the same associated sample sizes. 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, which is the default in all value-added applications, it is difficult to apply these methods to construct (2.7), given the observations $(y_1, \ldots, y_J)$ are no longer identically distributed.

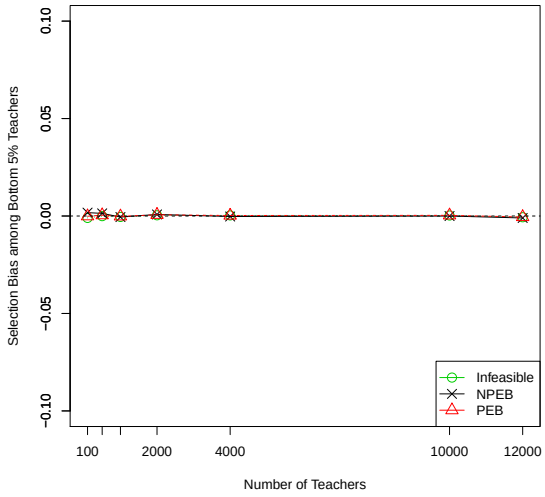
[72] See Koenker and Mizera (2014) for a monotonized version of the Brown and Greenshtein (2009) estimator.

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

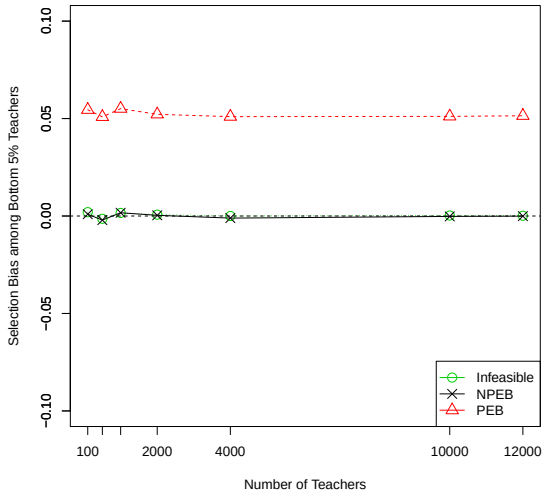
I Appendix Figures and Tables

Figure I.1: Simulation Performance in Terms of Selection Bias among Bottom 5% of Teachers for PEB and NPEB Estimators Relative to Infeasible Benchmark when Teacher Quality Follows:

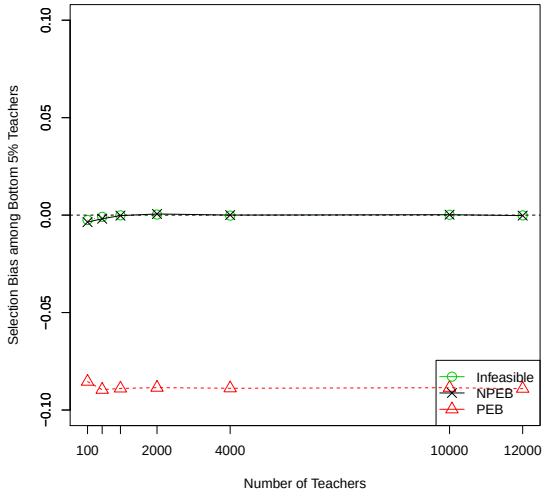
(a) Normal Distribution



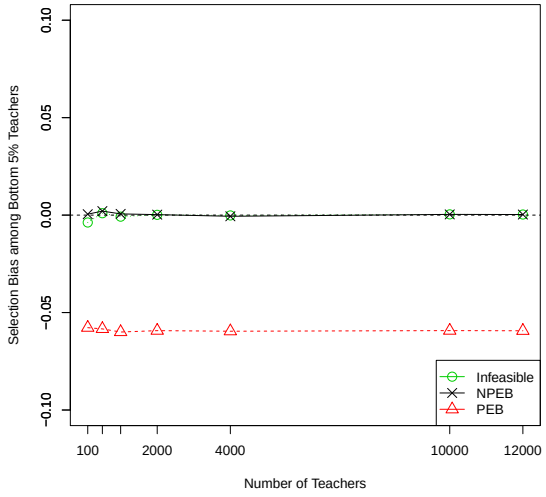
(b) Mixed Normal Distribution



(c) Skewed Normal Distribution

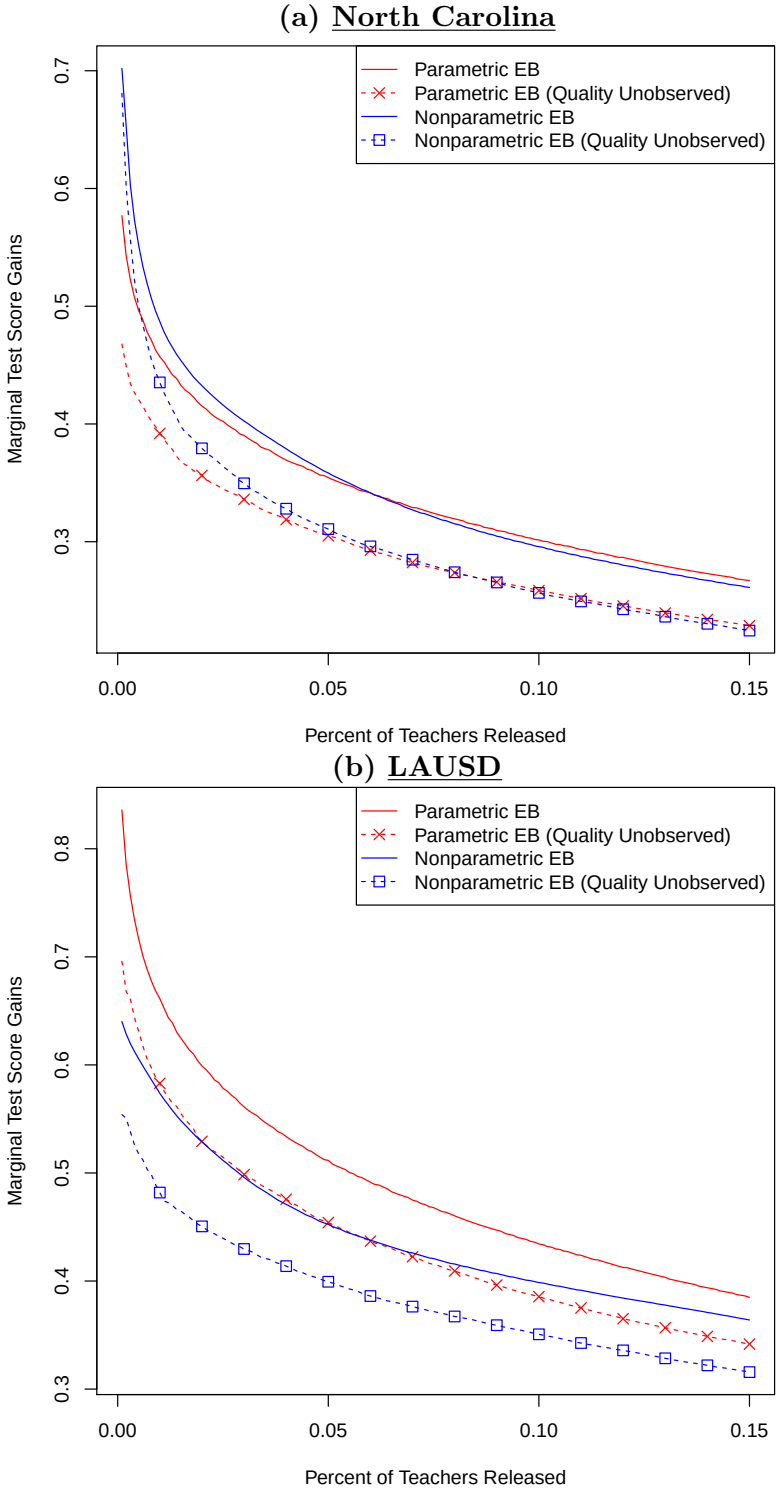


(d) Skewed Logistic Distribution



Notes: These figures assess the performance in terms of selection bias among bottom 5% teachers for the parametric (PEB) and our nonparametric (NPEB) estimators in finite sample relative to the infeasible benchmark for the four data generating processes we consider in our simulations. Each figure reports the selection bias among bottom 5% teachers of the PEB and NPEB estimators. We also show the selection bias for the infeasible estimator when the underlying distribution is known; the bias here is always close to zero. Figure 2(a) simulates the selection bias when the underlying distribution is normally distributed $F \sim \mathcal{N}(0, 0.08)$, while Figure 2(b) does so when the underlying distribution is mixed normal $F \sim 0.95\mathcal{N}(0, 0.03) + 0.025\mathcal{N}(-1, 0.03) + 0.025\mathcal{N}(1, 0.03)$. Figure 2(c) then skews the normal distribution to the right with location parameter -0.4 and shape parameter 5, while Figure 2(d) uses a skewed logistic with location parameter -0.5 and shape parameter 5. The parameters are chosen such that all four distributions are mean zero and have roughly the same variance. The x-axis reports number of teachers, each with a class size from the set {8, 16}. The simulations average results from 500 repetitions and set $\sigma_\epsilon^2 = 0.25$.

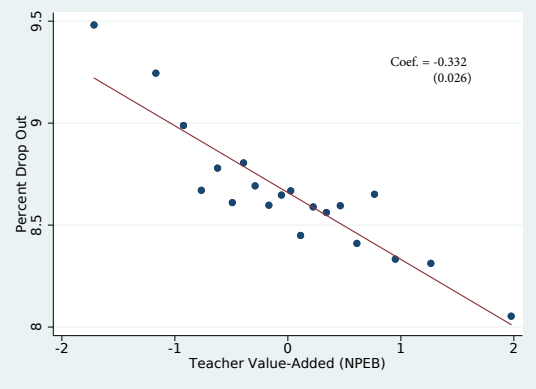
Figure I.2: Classroom Shocks Model: Test Scores Gains from Replacing Bottom q Percentile of Teachers when VA is Estimated



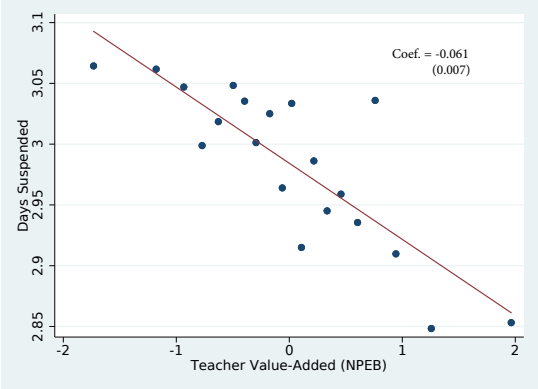
Notes: These figures show the average test score gains among affected students of a policy that releases bottom q% of teachers (replacing with mean-quality teachers) under the classroom shocks model presented in equation (8.1). The dashed lines represent the policy gains expected under the PEB and NPEB methodology when true teacher VA is estimated, while the solid lines indicate the policy gains if true teacher VA were observed. Policy gains when teacher VA is estimated 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.)

Figure I.3: Effects of Teacher Value-Added on Long-Run Outcomes (North Carolina)

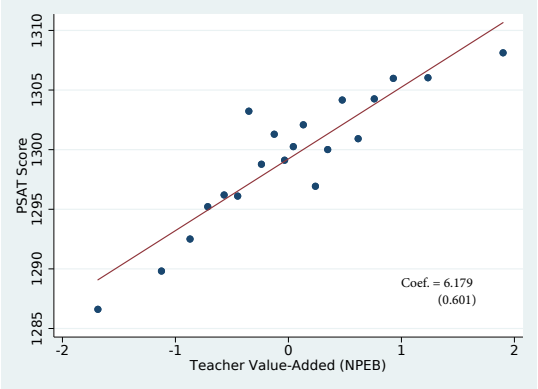
(a) High School Drop Out



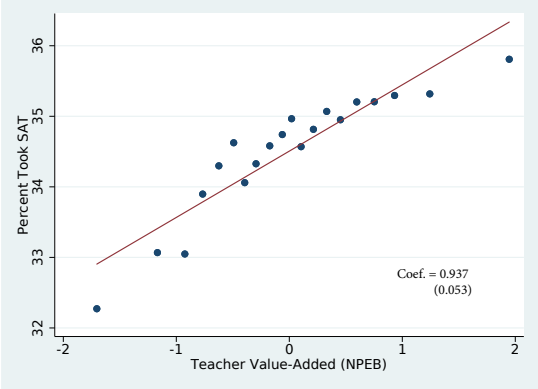
(b) Days Suspended



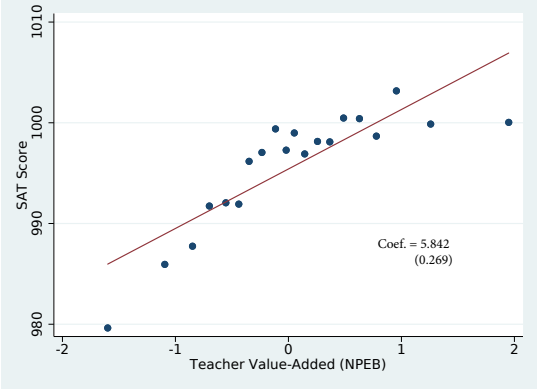
(c) PSAT Score



(d) Took SAT



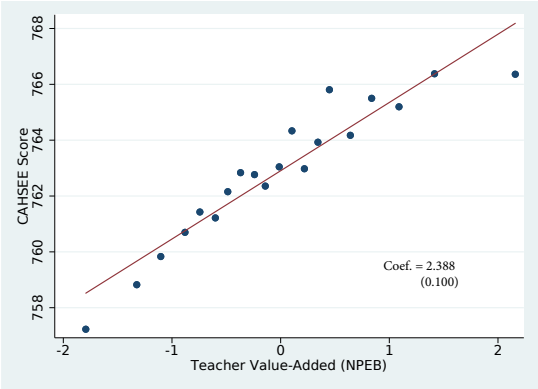
(e) SAT Score



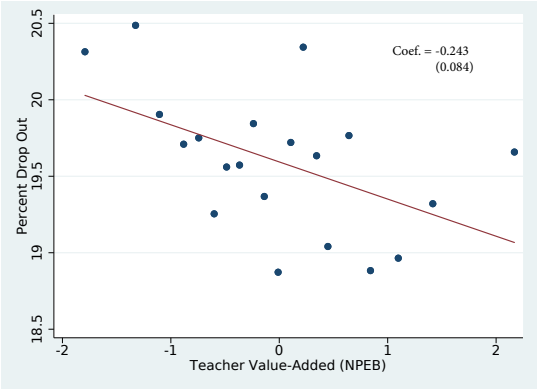
Notes: These figures link long-run outcomes to teacher value-added estimated using NPEB in North Carolina by comparing the long-run outcomes of students who were assigned to teachers with different value-added, controlling for a rich set of student characteristics. To do so, we follow the steps outlined in Chetty et al. (2014b): (i) residualize the long-run outcome as described in equation (D.1), (ii) divide our value-added estimates $\hat{m}_{jt}^{NPEB}$ into twenty equal-sized bins and plot the mean of the long-run outcome residuals in each bin against the bin mean of $\hat{m}_{jt}^{NPEB}$, and (iii) add back in the mean long-run outcome in the estimation sample to facilitate the interpretation of the scale. Coefficient estimates are reported in the figures with standard errors clustered at the school by cohort level in parentheses below. Coefficient estimates are the same as those reported for κ^{NPEB} in Table 5(a). Long-run outcomes are: high school drop out, total days suspended in middle and high school, PSAT scores, whether student takes the SAT, and SAT scores. For PSAT and SAT scores, we combine the mathematics and English components and take the values from the student's first attempt. (See Appendix C.3 for more details about the data construction.)

Figure I.4: Effects of Teacher Value-Added on Long-Run Outcomes (LAUSD)

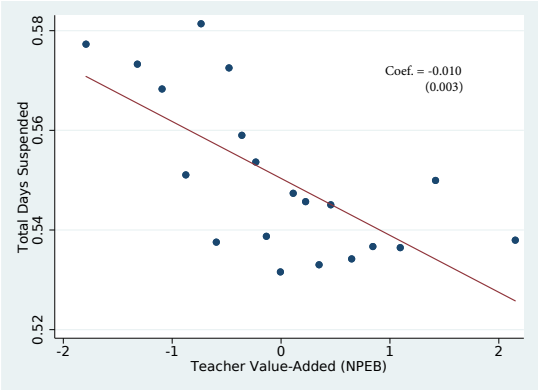
(a) Exit Exam Score (CAHSEE)



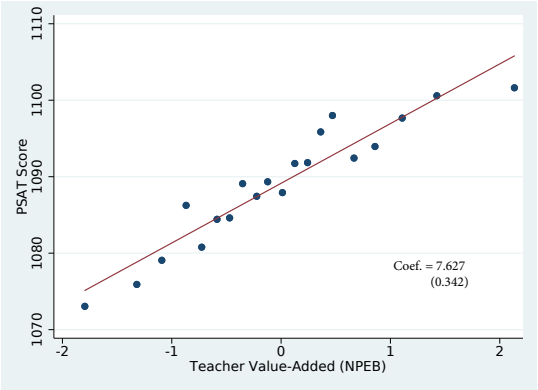
(b) High School Drop Out



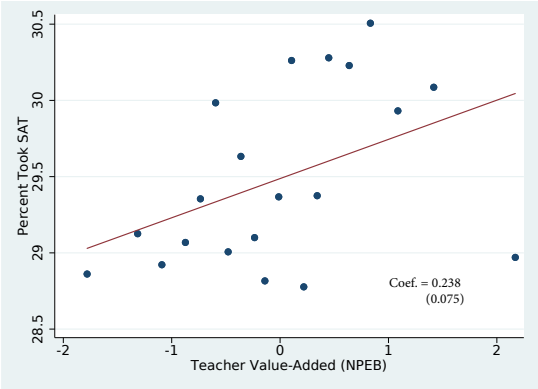
(c) Days Suspended



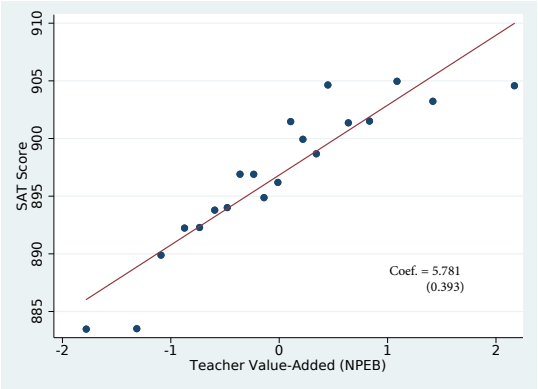
(d) PSAT Score



(e) Took SAT



(f) SAT Score



Notes: These figures link long-run outcomes to teacher value-added estimated using NPEB in LAUSD by comparing the long-run outcomes of students who were assigned to teachers with different value-added, controlling for a rich set of student characteristics. To do so, we follow the steps outlined in Chetty et al. (2014b): (i) residualize the long-run outcome as described in equation (D.1), (ii) divide our value-added estimates $\hat{m}_{jt}^{NPEB}$ into twenty equal-sized bins and plot the mean of the long-run outcome residuals in each bin against the bin mean of $\hat{m}_{jt}^{NPEB}$, and (iii) add back in the mean long-run outcome in the estimation sample to facilitate interpretation of the scale. Coefficient estimates are reported in the figures with standard errors clustered at the school by cohort level in parentheses below. Coefficient estimates are the same as those reported for κ^{NPEB} in Table 5(b). Long-run outcomes are: high school exit exam (CAHSEE) scores, high school drop out (vs. graduation), total days suspended in middle and high school, PSAT scores, whether student takes the SAT, and SAT scores. For CAHSEE, PSAT, and SAT scores, we combine the mathematics and English components and take the values from the student's first attempt. (See Appendix C.3 for more details about the data construction.)

Table I.1(a): Tail Selection Bias Simulation – *True Distribution is Normal*

	Homogeneous Class Sizes (Class Size of 16)				Heterogeneous Class Sizes (Class Size 8-16)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Bias in Top 5%	0.000	0.000	0.000	0.104	0.000	0.000	0.000	0.140
Bias in Bottom 5%	0.000	0.000	0.000	-0.104	0.000	0.000	0.000	-0.140

Table I.1(b): Tail Selection Bias Simulation – *True Distribution is Mixed Normal*

	Homogeneous Class Sizes (Class Size of 16)				Heterogeneous Class Sizes (Class Size 8-16)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Bias in Top 5%	0.000	0.000	-0.037	0.085	0.000	0.000	-0.052	0.124
Bias in Bottom 5%	0.000	0.000	0.037	-0.085	0.000	0.000	0.051	-0.125

Table I.1(c): Tail Selection Bias Simulation – *True Distribution is Skewed Normal*

	Homogeneous Class Sizes (Class Size of 16)				Heterogeneous Class Sizes (Class Size 8-16)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Bias in Top 5%	0.000	0.000	-0.042	-0.069	0.000	0.000	-0.056	0.104
Bias in Bottom 5%	0.000	0.000	-0.075	-0.150	0.000	0.000	-0.089	-0.207

Table I.1(d): Tail Selection Bias Simulation – *True Distribution is Skewed Logistic*

	Homogeneous Class Sizes (Class Size of 16)				Heterogeneous Class Sizes (Class Size 8-16)			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
Bias in Top 5%	0.000	0.001	-0.048	0.063	0.000	0.000	-0.064	0.095
Bias in Bottom 5%	0.000	0.000	-0.047	-0.124	0.000	0.000	-0.059	-0.178

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 selection bias when focusing on teachers in the tails (bottom and top five percent) of the VA distribution. The infeasible benchmark is the optimal empirical Bayes 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 $F \sim \mathcal{N}(0, 0.08)$. In Table 1(b), the underlying teacher VA distribution is mixed normal $F \sim 0.95\mathcal{N}(0, 0.03)+0.025\mathcal{N}(-1, 0.03)+0.025\mathcal{N}(1, 0.03)$. The normal and mixed normal distributions have the same first three moments (mean, variance, and skewness) for comparability. Table 1(c) then skews the normal distribution to the right with location parameter -0.4 and shape parameter 5, while Table 1(d) uses a skewed logistic with location parameter -0.5 and shape parameter 5. The parameters are chosen such that all four distributions are mean zero and have roughly the same variance. 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 for homogeneous class sizes (where every teacher has a class size of sixteen) and on the right side, for heterogeneous class sizes (where class sizes are drawn randomly from the set $\{8, 16\}$ with equal probability). These tables are an expansion on Tables 1(a)-1(d) which report the sum squared error of these simulations. Tables I.2(a)-I.2(d) further expand on these simulations by giving type I and II error rates for these simulations.

Table I.2(a): Teacher Ranking Simulation – *True Distribution is Normal*

	Homogeneous Class Sizes				Heterogeneous Class Sizes			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
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%	167.7	167.7	167.7	167.7	177.9	177.9	177.9	180.2
Top 5%	167.9	167.9	167.9	167.9	178.1	178.2	178.1	180.3

Table I.2(b): Teacher Ranking Simulation – *True Distribution is Mixed Normal*

	Homogeneous Class Sizes				Heterogeneous Class Sizes			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
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%	136.6	136.6	136.6	136.6	153.2	153.2	152.5	157.8
Top 5%	136.0	136.0	136.0	136.0	152.4	152.4	152.7	157.0

Table I.2(c): Teacher Ranking Simulation – *True Distribution is Skewed Normal*

	Homogeneous Class Sizes				Heterogeneous Class Sizes			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
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%	254.7	254.7	254.7	254.7	278.9	279.1	278.7	284.1
Top 5%	110.7	110.7	110.7	110.7	131.7	131.7	132.6	132.5

Table I.2(d): Teacher Ranking Simulation – *True Distribution is Skewed Logistic*

	Homogeneous Class Sizes				Heterogeneous Class Sizes			
	Infeasible	NPEB	PEB	FE	Infeasible	NPEB	PEB	FE
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%	203.7	203.7	203.7	203.7	230.8	230.8	230.1	234.6
Top 5%	99.3	99.3	99.3	99.3	118.9	119.0	119.6	119.8

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 type I and type II error in ranking bottom and top 5% teachers. Since type I error is the same as type II error by definition, we only report results for type I error. The infeasible benchmark is the optimal empirical Bayes 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 $F \sim \mathcal{N}(0, 0.08)$. In Table 1(b), the underlying teacher VA distribution is mixed normal $F \sim 0.95\mathcal{N}(0, 0.03) + 0.025\mathcal{N}(-1, 0.03) + 0.025\mathcal{N}(1, 0.03)$. The normal and mixed normal distributions have the same first three moments (mean, variance, and skewness) for comparability. Table 1(c) then skews the normal distribution to the right with location parameter -0.4 and shape parameter 5, while Table 1(d) uses a skewed logistic with location parameter -0.5 and shape parameter 5. The parameters are chosen such that all four distributions are mean zero and have roughly the same variance. The simulations average results from 500 repetitions with 10,000 individual teachers setting $\sigma_\epsilon^2 = 0.25$. Results are reported on the left side of each panel for homogeneous class sizes (where every teacher has a class size of sixteen) and on the right side, for heterogeneous class sizes (where class sizes are drawn randomly from the set $\{8, 16\}$ with equal probability). These tables are an expansion on Tables 1(a)-1(d) which report the sum squared error of these simulations. Tables I.1(a)-I.1(d) further expand on these simulations by giving selection bias in the tails (top and bottom 5%) of the distribution.

Table I.3: Quantiles of the Estimated NPMLE and Normal Distributions

Quantile:	0.003	0.01	0.05	0.1	0.2	0.3	0.5	0.7	0.8	0.9	0.95	0.99	0.997
Panel A. North Carolina													
NPMLE	-0.682	-0.454	-0.336	-0.230	-0.146	-0.093	0.024	0.091	0.149	0.274	0.338	0.490	0.657
Normal	-0.596	-0.505	-0.357	-0.278	-0.183	-0.114	0.000	0.114	0.183	0.278	0.357	0.505	0.596
Panel B. LAUSD													
NPMLE	-0.669	-0.560	-0.435	-0.344	-0.207	-0.138	-0.020	0.143	0.224	0.379	0.465	0.670	0.847
Normal	-0.859	-0.727	-0.514	-0.401	-0.263	-0.164	0.000	0.164	0.263	0.401	0.514	0.727	0.859

Notes: This table shows the teacher VA cutoff values for given quantiles of the estimated NPMLE and normal distribution in both the North Carolina and LAUSD data. For example, the values in the first two rows of each panel at 0.05 represent the VA of the bottom 5% teacher according to the VA distribution estimated via NPMLE and when normality is assumed, respectively. These cutoffs values are the same as the visual representation of the quantiles in Figure 4.

Table I.4: Test Scores Gains from Releasing Bottom q Percentile Teachers
(True VA Observed)

% Teachers Released (q)	North Carolina Data		LAUSD Data	
	Test Gain under F (NPEB)	Test Gain under Normal (PEB)	Test Gain under F (NPEB)	Test Gain under Normal (PEB)
	(1)	(2)	(3)	(4)
1	0.577	0.579	0.662	0.833
	(0.008)	(0.002)	(0.012)	(0.006)
3	0.474	0.492	0.572	0.709
	(0.004)	(0.002)	(0.007)	(0.005)
5	**0.424**	**0.447**	**0.523**	**0.645**
	(0.004)	(0.002)	(0.006)	(0.004)
7	0.389	0.416	0.489	0.600
	(0.003)	(0.002)	(0.005)	(0.004)
9	0.362	0.391	0.462	0.0564
	(0.003)	(0.002)	(0.004)	(0.004)

Notes: Table I.4 displays the estimated gains in mathematics scores among affected students 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 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 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. The bolded line indicates the widely-analyzed 'release the bottom five percent of teachers' policy. Standard errors are calculated using the bootstrap as described in Appendix E. These tables are analogous to Table 4 which reports policy gains when true teacher quality is *unobserved* by the policymaker.

Table I.5(a): Long-Run Gains from Releasing Bottom 5% of Teachers – True VA Observed (North Carolina)

Long-Run Outcome:	Percent Drop Out (1)	Days Suspended (2)	PSAT Score (3)	Percent Took SAT (4)	SAT Score (5)	Exit Exam Score (6)
Sample Mean	8.63	2.98	1299.8	34.6	996.3	-
Panel A. Parametric Empirical Bayes						
Benefit ($\hat{\kappa}^{PEB}$)	-0.36	-0.065	6.63	1.01	6.33	-
Average Change in VA of Released Teachers (Δm_{σ}^{PEB})	2.06	2.06	2.06	2.06	2.06 -	
Gain of Releasing Bottom 5% (G^{PEB})	-0.74	-0.135	13.68	2.07	13.06	-
Panel B. Nonparametric Empirical Bayes						
Benefit ($\hat{\kappa}^{NPEB}$)	-0.33	-0.061	6.18	0.94	5.84	-
Average Change in VA of Released Teachers (Δm_{σ}^{NPEB})	2.11	2.11	2.11	2.11	2.11	-
Gain of Releasing Bottom 5% (G^{NPEB})	-0.70	-0.129	13.07	1.98	12.36	-
Overestimation of Parametric EB (%)	4.8	4.6	4.7	4.7	5.7	-

Table I.5(b): Long-Run Gains from Releasing Bottom 5% of Teachers – True VA Observed (LAUSD)

Long-Run Outcome:	Percent Drop Out (1)	Days Suspended (2)	PSAT Score (3)	Percent Took SAT (4)	SAT Score (5)	Exit Exam Score (6)
Sample Mean	19.6	0.55	1089.7	29.5	897.3	763.0
Panel A. Parametric Empirical Bayes						
Benefit ($\hat{\kappa}^{PEB}$)	-0.28	-0.012	8.74	0.27	6.65	2.74
Average Change in VA of Released Teachers (Δm_{σ}^{PEB})	2.06	2.06	2.06	2.06	2.06	2.06
Gain of Releasing Bottom 5% (G^{PEB})	-0.59	-0.024	18.03	0.57	13.73	5.65
Panel B. Nonparametric Empirical Bayes						
Benefit ($\hat{\kappa}^{NPEB}$)	-0.24	-0.010	7.63	0.24	5.78	2.39
Average Change in VA of Released Teachers (Δm_{σ}^{NPEB})	1.92	1.92	1.92	1.92	1.92	1.92
Gain of Releasing Bottom 5% (G^{NPEB})	-0.47	-0.019	14.64	0.46	11.10	4.58
Overestimation of Parametric EB (%)	25.7	24.2	23.1	23.5	23.7	23.3

Notes: Tables I.5(a) and I.5(b) show – using the North Carolina and LAUSD data, respectively – the estimated gains in terms of various long-run outcomes of a policy that releases the bottom 5% of teachers and replaces them with mean quality teachers when true teacher quality is *observed* by the policymaker. The 'Benefit' row in each panel represents the increase in the long-run outcome associated with having a teacher whose VA is one standard deviation higher, as described by equation (D.3); this benefit for NPEB is shown graphically in Figures I.3 and I.4 for North Carolina and LAUSD, respectively. The average value-added of bottom 5% teachers is then calculated, and multiplied by the benefit to give the policy effect of releasing bottom 5% teachers according to VA (given by third row of each panel). The final row of each table, 'Overestimation of Parametric EB,' gives the overestimation of policy gains in terms of the long-run outcome from utilizing PEB rather than NPEB. Long-run outcomes are: high school drop out, total days suspended in middle and high school, PSAT scores, whether student takes the SAT, and SAT scores, and exit exam scores (only available for LAUSD). For PSAT and SAT scores, we combine the mathematics and English components and take the values from the student's first attempt. (See Appendix C.3 for more details about data construction.) Tables 5(a) and 5(b) repeat the exercise here when true teacher quality is unobserved to the policymaker and so teacher releases are based on estimated (rather than true) value-added: the results from doing so are virtually identical.