# A New Method for Estimating Teacher Value-Added[*]

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

September 18, 2019

## Abstract

This paper proposes a new methodology for estimating teacher value-added. Rather than imposing a normality assumption on unobserved teacher quality (as in the standard empirical Bayes approach), our nonparametric estimator permits the underlying distribution to be estimated directly and in a computationally feasible way. We implement the approach using two separate large-scale administrative data sets, finding that our estimated teacher value-added distributions depart from normality and differ from each other. We then draw out the implications of our method for policies that are based on teacher value-added. First, considering a widely discussed policy that releases teachers in the bottom five percent of the value-added distribution, we compare predicted test score gains using our approach with those using parametric empirical Bayes. The parametric method predicts similar policy gains in one data set while overestimating the gains in the other by around 30 percent. More generally, we simulate the aggregate test score effects of policies that release any given percentage of teachers from the bottom of the value-added distributions under the two approaches, and also policies that reassign teachers at the top of the distribution. The results highlight the benefit of using a nonparametric approach, given that the underlying (unobserved) distribution of value-added is likely to be context-specific.

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

[*] We would like to thank Roger Koenker for helpful discussions, and Joe Altonji, Raj Chetty, Michael Dinerstein, John Friedman, Chris Taber and seminar participants at the NBER Summer Institute 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,[1] researchers have proposed influential fixed effects methods as a means to capture a teacher's overall quality, taking advantage of large-scale matched student-teacher data sets that are increasingly accessible – see Rockoff (2004) and Rivkin, Hanushek, and Kain (2005) for pioneering studies. In turn, fixed effects methods have prompted the development of teacher value-added (VA) estimators for measuring the impact of teachers that are transparent and easy to implement.[2] Given the appeal of such VA estimators, teacher value-added estimates now feature ever more widely in the policy sphere, particularly in consequential teacher retention, promotion and pay decisions. Indeed, by the end of 2017, fully thirty nine states required value-added measures to be incorporated into teacher evaluation scores (as one indicator of this phenomenon).

The use of value-added methods in high-stakes decision-making raises important challenges. Not least, such methods need to be able to identify teacher quality on the basis of relatively few teacher-year observations. The standard approach to this issue involves using empirical Bayes methods to reduce measurement error in value-added estimates, 'shrinking' less reliable estimates back toward the mean (Kane and Staiger, 2008; Kane, Rockoff, and Staiger, 2008; Jacob and Lefgren, 2008; Harris and Sass, 2014; Chetty, Friedman, and Rockoff, 2014a,b). In order to apply these methods, papers estimating teacher value-added have typically used the best linear prediction to estimate teacher quality. This parametric empirical Bayes approach, first proposed by Morris (1983), is attractive both for its analytic convenience and given that it is the optimal Bayes rule for estimating unobserved teacher quality when underlying teacher quality is normally distributed.

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

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

[2] See Koedel, Mihaly, and Rockoff (2015) for an up-to-date review.

issues.

The central contribution of our paper is to propose a feasible new methodology for estimating teacher value-added that does not impose any parametric assumptions on the unobserved heterogeneity in teacher quality. Our approach uses the nonparametric empirical Bayes ('NPEB') estimator due to Robbins (1956). In this, we follow the prior literature that implements empirical Bayes ('EB'), but rather than assuming teacher quality to be normally distributed, we *estimate* the underlying teacher quality distribution using a nonparametric maximum likelihood estimator.

Our proposed estimator is optimal in the sense of minimizing the variance of individual teacher quality estimates *regardless* of the true underlying distribution of unobserved teacher quality.[3] Starting with a noisy measure of true teacher quality, the teacher quality distribution can be identified nonparametrically.[4] In addition, the approach allows the teacher quality estimator to depend on different classroom sizes and other sources of student achievement heteroscedasticity in a nonlinear way, thereby extending previous methods. Pioneering computational advances by Koenker and Mizera (2014) have only recently made the implementation of Robbins' idea feasible in large-scale empirical applications.[5] We leverage those in the current study, showing first that the NPEB approach performs very well in Monte-Carlo simulations, regardless of the true underlying distribution; this is in marked contrast to the standard EB approach, as the simulations demonstrate.

To apply the methodology using observational data, we estimate teacher value-added in two separate large-scale administrative data sets: one covering the entire state of North Carolina and the other, a large urban school district in the west of the United States. Our estimated teacher value-added distributions differ both from the normal distribution assumed in the prior literature and from each other. In North Carolina, our estimated teacher distribution has a relatively similar shape to the normal, although with fatter tails; in the large urban school district, our estimated teacher distribution is skewed, with a much thinner left than right tail.

Given the deviations from normality in both settings, we then evaluate the policy relevance of our methodology. One advantage of our approach relative to the standard EB method is that

[3] To be precise, it minimizes the Bayes risk under $\mathcal{L}_2$ loss (defined below).

[4] We provide a deconvolution proof in the case of teacher value-added below. In the empirics, we make the assumption that the noise is independent of underlying teacher quality and has a known limiting distribution, assumed to be normal with individual teacher-specific variances. (As justification, the noise measure comes from a classroom average of test scores, and hence a central limit theorem applies.) Then further assuming that unobserved teacher quality *also* follows a normal distribution involves an over-parameterization.

[5] Other applications in economics include analyses of earnings dynamics (see Gu and Koenker (2017b)).

it should improve our ability to predict the impact of teacher quality reforms on students' future outcomes (to the extent that the normality assumption is misplaced). Here we focus on a teacher quality reform that has received considerable attention: releasing teachers in the bottom five percent of the estimated teacher value-added distribution, as proposed by Hanushek (2009, 2011) and evaluated in Chetty, Friedman, and Rockoff (2014b).

To assess whether our approach alters the estimated policy effects in an appreciable way, we compare the predicted test score gains of students when applying it with an approach that imposes normality on the underlying teacher quality distribution. Specifically, after observing each teacher for three years, we suppose the bottom five percent of teachers (based on estimated value-added) are released and replaced by teachers of average quality. In North Carolina, we find only minor differences between the two methods: the parametric empirical Bayes 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 large urban school district leads to large differences in the estimated policy benefit, with parametric empirical Bayes overstating test score gains of the policy by around 30 percent – a substantial amount.

At a general level, our methodology offers policymakers a means to better understand the benefits of implementing the same reform in different environments, given its data-driven emphasis. The policy application we present makes clear that the benefit of a reform releasing low value-added teachers can be substantially overestimated when the normality assumption is invoked – in one data set, at least. This points up the plausible notion that the true underlying distribution of teachers is likely to be context-specific, and so estimated policy gains may well differ substantially from the true policy gains in some settings (such as the large urban school district we analyze) when normality is invoked. Further, given the computational feasibility of our approach, analytical convenience need no longer weigh on the side of assuming normality. We describe other education applications below.

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

# 2 Methodology

This section sets out our methodology for estimating teacher value-added. We start with a standard model of student achievement, then place our method in the context of the main existing approaches used in the literature.

## 2.1 Student Test Scores and the Contribution of Teachers

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

$$A^*_{ijt} = X^\top_{ijt}\beta + \alpha_j + \epsilon_{ijt}, \quad i = 1, 2, \ldots, n_{jt}, \tag{2.1}$$

where $A^*_{ijt}$ is the student's test score, $X_{ijt}$ are observed covariates that characterize the student's demographics, past academic performance, family background and also the teacher (including her experience). Our parameter of interest is the time-invariant teacher's contribution (or simply 'valued-added'), $\alpha_j$. We assume that teachers are each assigned to one class per year (with class size $n_{jt}$) and that, conditional on $X_{ijt}$, the assignment is as good as random; the error term $\epsilon_{ijt}$ is assumed to be iid with variance $\sigma^2_\epsilon$.

The standard approach to estimating teacher value-added involves two steps. First, a regression is used to purge the effects of observed covariates from $A^*_{ijt}$. What is left is a 'noised' measure of the teacher's contribution, denoted

$$A_{ijt} = A^*_{ijt} - X^\top_{ijt}\hat{\beta} = \alpha_j + \epsilon_{ijt} + X^\top_{ijt}(\hat{\beta} - \beta) \approx \alpha_j + \epsilon_{ijt}\,. \tag{2.2}$$

Second, averaging over each classroom, applying the central limit theorem, and using the fact that $\hat{\beta}$ converges to $\beta$ in probability, we have

$$y_{jt} \equiv \frac{1}{n_{jt}} \sum_{i=1}^{n_{jt}} A_{ijt} \overset{d}{\to} \mathcal{N}(\alpha_j, \frac{\sigma^2_\epsilon}{n_{jt}}). \tag{2.3}$$

From here, there are several ways to proceed to estimate the value-added measure, $\alpha_j$. We will assess different estimators by their distance from the true value-added quantity. A commonly-used distance measure is the so-called $\mathcal{L}_2$ loss, which we will use, computed as the expected value of $L(\hat{\delta}, \alpha) \equiv (\hat{\delta} - \alpha)^2$, where $\hat{\delta}$ is some estimator of true value-added, $\alpha$.

## 2.2 The Fixed Effect Estimator

Given the result in (2.3), we can construct the maximum likelihood estimator (sometimes referred to as the fixed effect estimator) for the unobserved $\alpha_j$ as

$$\bar{v}_j = \sum_t h_{jt} y_{jt} / \sum_t h_{jt} = \sum_t n_{jt} y_{jt} / \sum_t n_{jt} \ ,$$

where the weight $h_{jt} \equiv n_{jt}/\sigma_\epsilon^2$. Applying the result in (2.3) again, the fixed effect estimator converges to the following distribution:

$$\bar{v}_j \xrightarrow{d} \mathcal{N}(\alpha_j, \sigma_\epsilon^2 / \sum_t n_{jt}). \tag{2.4}$$

If the total sample $\sum_t n_{jt} \to \infty$, then the maximum likelihood estimator $\bar{v}_j$ converges to the true teacher value-added $\alpha_j$ in probability, and so is a consistent estimator for the desired object.

In practice, the value-added literature does not use the fixed effect estimator, primarily because of finite sample considerations.[6] Instead, the current state-of-art estimator for value-added – the parametric empirical Bayes estimator introduced first by Kane and Staiger (2008) and further developed by Chetty, Friedman, and Rockoff (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. This leads to the parametric empirical Bayes estimator, which is the sample analogue of $\hat{\alpha}_j$ from Bayes rule, minimizing Bayes risk under the average loss $\frac{1}{J}\sum_{j=1}^{J}(\hat{\alpha}_j - \alpha_j)^2$.

## 2.3 The Parametric Empirical Bayes Estimator

Given (2.4) and the assumption that the value-added for all teachers is an independent and identically distributed draw from a normal distribution with mean zero and variance $\sigma_\alpha^2$, the minimizer

[6] As a consequence of these, the fixed-effect estimator $\bar{v}_j$ is a noisy estimator, especially for teachers just beginning their careers.

of the Bayes risk, given by $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}(\delta_j - \alpha_j)^2]$, takes the following form:

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

This is the teacher value-added estimator in Kane and Staiger (2008) and Chetty *et al.* (2011). Several remarks about the estimator in (2.5) are due:

1. When $\alpha_j \sim_{iid} \mathcal{N}(0, \sigma_\alpha^2)$, the posterior distribution of $\alpha_j$ conditional on observing the teacher's performance $\bar{v}_j$ also follows a normal distribution, given as $\alpha_j \mid \bar{v}_j \sim \mathcal{N}(\bar{v}_j \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2/\sum_t n_{jt}}, \frac{\sigma_\alpha^2 \sigma_\epsilon^2/\sum_t n_{jt}}{\sigma_\alpha^2 + \sigma_\epsilon^2/\sum_t n_{jt}})$ and the posterior mean of $\alpha_j$ given the observation $\bar{v}_j$ is the best linear predictor of $\alpha_j$.

2. The shrinkage factor, $\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2/\sum_t n_{jt}}$ is always smaller than 1, which implies that the Bayes estimator $\delta_j$ shrinks the fixed effect estimator $\bar{v}_j$ towards zero.

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

4. There is built-in symmetry in the Bayes estimator in the sense that, for individual teachers who have the same total sample size, the amount of shrinkage imposed on $\bar{v}_j$ only depends on its absolute value. Thus teachers with very large $\bar{v}_j$ (for example, teachers in the right tail with large positive fixed effects) and with very small $\bar{v}_j$ (for example, left-tail teachers) are shrunk towards zero by the same magnitude as long as their effective sample sizes are the same.

5. The estimator $\delta_j$ is infeasible since it involves unknown parameters $(\sigma_\alpha^2, \sigma_\epsilon^2)$. The empirical counterpart of $\delta_j$, the empirical parametric Bayes estimator, replaces these unknown parameters with their consistent estimates either through maximum likelihood or method of moments.

These observations serve to highlight two features of the parametric empirical Bayes estimator: First, by construction, the estimator $\delta_j$ is linear in $\bar{v}_j$; hence it scales the fixed effect estimator in the same symmetric fashion towards zero for all teachers who have the same overall sample size $n_j$.

Second, for teachers who have very large $n_j$ relative to $\sigma_\epsilon^2$, the empirical Bayes estimator is almost the same as the fixed effect $\bar{v}_j$. This confirms the intuition from above that if $n_j$ is very large, the fixed effect will provide an accurate estimator for the true value-added, so the shrinkage estimator $\delta_j$ leaves it relatively unmodified. However, teachers with a smaller total sample size (for example new teachers) receive the most shrinkage towards zero.

The empirical parametric Bayes estimator has a very simple linear form and is easy to compute, helping to account for its popularity in the literature. However, it relies crucially on the parametric assumption that true teacher value-added $\alpha_j$ follows a normal distribution. If this assumption is misplaced, the quality of the shrinkage estimator may deteriorate significantly. This raises the possibility that one might be able to find an alternative estimator that has a much smaller Bayes risk.

We show next that given (2.4), the distribution of teacher value-added is in fact nonparametrically identified (Theorem 1 below). This is the generalized deconvolution problem which has been considered in Delaigle and Meister (2008). It implies that the data contain enough information about the distribution of the true value-added measure, and the normality assumption involves an over-parameterization.

We then introduce the nonparametric empirical Bayes (or 'NPEB') estimator for teacher value-added (see Theorem 2). Our Monte Carlo simulations demonstrate that with many relevant data generating processes, the new estimator performs significantly better than the parametric empirical Bayes estimator in terms of estimation risk. We also show that the NPEB estimator for teacher value-added is empirically relevant when applied to actual large-scale administrative data sets in education.

## 2.4 The Nonparametric Empirical Bayes Estimator

**Theorem 1** *Given the model in which $\bar{v}_j = \alpha_j + \nu_j$ for $j = 1, \ldots, J$ and $t = 1, \ldots, T$, with $\nu_j \sim \mathcal{N}(0, \sigma_\epsilon^2/n_j)$, if $\alpha_j$ is independent of $\nu_j$, and $\alpha_j$ follows some probability distribution $F$, then $F$ is nonparametrically identified.*

**Proof.** Under the assumption that $\alpha_j$ and $\nu_j$ are independent random variables, we have for any

$t \in \mathbb{R}$,

$$\begin{aligned}
\varphi_{\bar{v}_j}(t) &= \int e^{it\bar{v}_j} \int \frac{1}{\sqrt{2\pi\sigma_\epsilon^2/n_j}} e^{-\frac{(\bar{v}_j-\alpha)^2}{2\sigma_\epsilon^2/n_j}} dF(\alpha) d\bar{v}_j \\
&= \int e^{itz} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2/n_j}} e^{-\frac{z^2}{2\sigma_\epsilon^2/n_j}} dz \int e^{it\alpha} dF(\alpha) \\
&= e^{-\sigma_\epsilon^2 t^2/2n_j} \varphi_\alpha(t)
\end{aligned}$$

where $\varphi_X(t)$ is the characteristic function of random variable $X$. Since we observe $\bar{v}_j$, the characteristic function $\varphi_\alpha(t)$ is identified from the data for all $t \in \mathbb{R}$. Given the one-to-one mapping from the characteristic function to the distribution function of a random variable, the distribution $F$ is nonparametrically identified. ■

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

$$\tilde{\alpha}_j = \bar{v}_j + \frac{\sigma_\epsilon^2}{\sum_t n_{jt}} \frac{\partial}{\partial \bar{v}} \log g_j(\bar{v})|_{\bar{v}=\bar{v}_j} ,$$

*with $g_j(\bar{v}) = \int \phi(\bar{v}; \alpha, \sigma_\epsilon^2/n_j) dF(\alpha)$, where $\phi(\cdot, a, b)$ is the normal density with mean $a$ and variance $b$.*

**Proof.** The minimizer of the Bayes risk under $\mathcal{L}_2$ loss is nothing but the posterior mean of $\alpha$ conditional on $\bar{v}_j$. Therefore,

$$\begin{aligned}
\mathbb{E}(\alpha|\bar{v}_j) &= \int \alpha \frac{1}{\sqrt{2\pi\sigma_\epsilon^2/n_j}} e^{-(\bar{v}_j-\alpha)/2\sigma_\epsilon^2/n_j} dF(\alpha) / \int \frac{1}{\sqrt{2\pi\sigma_\epsilon^2/n_j}} e^{-(\bar{v}_j-\alpha)/2\sigma_\epsilon^2/n_j} dF(\alpha) \\
&= \int \alpha e^{\frac{\alpha\bar{v}_j}{\sigma_\epsilon^2/n_j}} e^{-\frac{\alpha^2}{2\sigma_\epsilon^2/n_j}} dF(\alpha) / \int e^{\frac{\alpha\bar{v}_j}{\sigma_\epsilon^2/n_j}} e^{-\frac{\alpha^2}{2\sigma_\epsilon^2/n_j}} dF(\alpha) \\
&= \frac{\sigma_\epsilon^2}{n_j} \frac{\partial}{\partial \bar{v}} \log \Big( \int e^{\frac{\alpha\bar{v}}{\sigma_\epsilon^2/n_j}} e^{-\frac{\alpha^2}{2\sigma_\epsilon^2/n_j}} dF(\alpha) \Big)|_{\bar{v}=\bar{v}_j} \\
&= \frac{\sigma_\epsilon^2}{n_j} \frac{\partial}{\partial \bar{v}} \log \Big( g_j(\bar{v}) / \Big( \frac{1}{\sqrt{2\pi\sigma_\epsilon^2/n_j}} e^{-\frac{\bar{v}^2}{2\sigma_\epsilon^2/n_j}} \Big) \Big)|_{\bar{v}=\bar{v}_j} \\
&= \bar{v}_j + \frac{\sigma_\epsilon^2}{n_j} \frac{\partial}{\partial \bar{v}} \log g_j(\bar{v})|_{\bar{v}=\bar{v}_j}
\end{aligned}$$

■

Theorem 2 is known as the Tweedie formula (see Robbins (1956) and Efron (2011)). We noted that the parametric Bayes estimator is a special case of Theorem 2. That is, when $F = \mathcal{N}(0, \sigma_\alpha^2)$, we have $g_j(\bar{v}) = \phi(\bar{v}; 0, \sigma_\alpha^2 + \sigma_\epsilon^2/n_j)$, $\frac{\partial}{\partial \bar{v}} \log g_j(\bar{v})|_{\bar{v}=\bar{v}_j} = -\frac{\bar{v}_j}{\sigma_\alpha^2+\sigma_\epsilon^2/n_j}$ and $\tilde{\alpha}_j = \delta_j = \bar{v}_j \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2/n_j}$.

For any distribution $F$ other than the normal distribution, the quantity $\frac{\partial}{\partial \bar{v}} \log g_j(\bar{v})|_{\bar{v}=\bar{v}_j}$ in Theorem 2 introduces a non-linearity into the shrinkage rule with respect to $\bar{v}_j$. The rule still retains the feature that if $\sigma_\epsilon^2/n_j$ is very small, then the nonparametric estimator $\tilde{\alpha}_j$ will not deviate much from the fixed effect estimator $\bar{v}_j$. However, unlike the parametric Bayes estimator, individuals with larger variances will not necessarily shrink the most towards zero.

This feature turns out to be empirically relevant: suppose a new teacher $j$ who has a small total sample size $n_j$ relative to $\sigma_\epsilon^2$ happens to have a relatively large $\bar{v}_j$ (for example, a teacher who performs very promisingly in her early years in the school system). Under the parametric shrinkage method, this teacher will be severely discounted since her value-added measure will shrink significantly towards zero simply because the normal distribution, which dictates that there must be a thin tail deems it unlikely for this teacher to have a large value-added $\alpha_j$, since the large $\bar{v}_j$ arises purely by chance due to the associated large variance. In contrast, under the nonparametric shrinkage rule, depending on the features of the distribution of true value-added, her value-added estimate may remain very close to her observed $\bar{v}_j$.[7]

**Example:** We provide a simple example, in line with this thought experiment. Suppose the distribution $F$ takes the following form:

$$F = 0.98\mathcal{N}(0, \theta_1) + 0.01\mathcal{N}(-1, \theta_2) + 0.01\mathcal{N}(1, \theta_3)\,. \tag{2.6}$$

This distribution has the feature that at both tails, there is a small probability mass concentrated around the values $-1$ and 1, while the majority of the probability mass follows a normal distribution centered at zero. We consider the Bayes estimator under two different assumptions regarding the true teacher value-added distribution. In the first case, assume $\alpha_j \sim \mathcal{N}(0, 0.05)$; looking ahead, this coincides with the empirical Bayes estimates obtained using the North Carolina data

[7] A similar thought experiment can be conducted for a teacher who shows very poor observed performance (that is, who has a large negative $\bar{v}_j$) and has a relatively small total sample size. Under the parametric shrinkage method, this teacher would be regarded as more similar to the mean quality teacher, while under the nonparametric method, this might not be the case.

on mathematics scores. In the second case, the true teacher value-added has distribution $F$, where we calibrate the above parameters $(\theta_1, \theta_2, \theta_3)$ such that the distribution $F$ has the same mean and variance as in the first case. In both, we set $\sigma_\epsilon^2 = 0.25$ and $n_j = 20$. (The value of the variance of $\epsilon$ is roughly the same as in the North Carolina mathematics score data.)

The density functions for these two cases are plotted in Figure 1(a). Figure 1(b) then compares the Bayes estimator for a hypothetical teacher with a total sample size equal to 20 and a fixed effect estimator that could take any value in the range of $[-1.5, 1.5]$. The black line in the figure is the Bayes estimator evaluated at different values of the fixed effect in this range. As expected, the greatest shrinkage occurs when this teacher has either very large or very small fixed effect estimates. In contrast, the dashed line corresponds to the Bayes estimator when the true teacher value-added distribution is not normal. The shrinkage behaves very similarly to the black curve in the middle range of the fixed effect estimate; however, the two curves depart at both tails. Were a teacher to have a relatively large or small fixed effect estimate, her value-added estimate would not be shrunk towards zero (as in the parametric Bayes case), but rather would stay close to her observed performance.

The above example illustrates that as the distribution $F$ deviates from the normal distribution, the nonparametric Bayes estimator might change significantly. But unless we know the distribution $F$, the nonparametric Bayes estimator $\tilde{\alpha}_j$ is infeasible in practice. We take an approach that shares the same spirit as the parametric empirical Bayes method where the unknown parameters in the normal distribution for $\alpha_j$ are estimated from the data directly, but instead estimating the distribution $F$ from the data, obtaining a nonparametric empirical Bayes estimator.

## 3 Estimation

This section describes the maximum likelihood approach we use to estimate teacher value-added nonparametrically.

### 3.1 Nonparametric Maximum Likelihood Estimation of the Distribution $F$

Our approach to recovering the teacher value-added distribution $F$ using nonparametric maximum likelihood estimation is based on methods proposed in Jiang and Zhang (2009) and Gu and

Koenker (2017b). Those papers propose a general estimation method for unobserved heterogeneity in cross-sectional and longitudinal data settings without imposing any parametric assumptions on the unobserved heterogeneity, drawing on the seminal contribution by Robbins (1956). The methodology fits many contemporary 'Big Data' applications, and there has been a recent revival in using these methods for large-scale inference in the statistics literature.[8] The teacher value-added application fits well into this framework, as teacher quality can be thought of as unobserved heterogeneity in the test score model that accounts for variation in test scores unexplained after controlling for all observed heterogeneity through $X_{it}$.[9]

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

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

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

Kiefer and Wolfowitz (1956) established consistency of the NPMLE for the mixing distribution $F$. Yet it was not until the appearance of the EM algorithm by Laird (1978) that a viable computational strategy for the estimator was available. The EM algorithm has remained the standard approach for its computation ever since.[10] However, EM has notoriously slow convergence in nonparametric EB applications, especially with large data sets, and this fact has seriously inhibited the use of the NPMLE. Koenker and Mizera (2014) have recently proposed an alternative computational method for the NPMLE that circumvents these issues. For a broad class of mixture problems, the Kiefer-Wolfowitz estimator can be formulated as a convex optimization problem

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

[9] The nonparametric empirical Bayes method has also been applied in other contexts in economics. For instance, Gu and Koenker (2017b) apply the methodology to earning dynamics, Gu and Koenker (2017a) use it to predict baseball batting averages, Gu and Shen (2017) analyze treatment effect heterogeneity, and Abadie and Kasy (forthcoming) estimate location effects on intergenerational mobility.

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

and solved efficiently by modern interior point methods. Quicker, more accurate computation of the NPMLE opens the way to a much wider range of applications of the method for models with heterogeneity. In particular, the large-scale data involved in teacher value-added estimation make it likely to benefit from the scalability of the new computational method.

### 3.2 The Plug-in Nonparametric Empirical Bayes Estimator for Value-added

With the NPMLE $\hat{F}$, we can construct the nonparametric empirical Bayes teacher quality estimator as:

$$\hat{\alpha}_j = \frac{\int \alpha \varphi_j(\bar{v}_j - \alpha) d\hat{F}(\alpha)}{\int \varphi_j(\bar{v}_j - \alpha) d\hat{F}(\alpha)}.$$

The newly proposed estimator $\hat{\alpha}_j$ has the potential to improve on the linear empirical Bayes estimator (in the sense of having a smaller average squared error) when the underlying distribution $F$ cannot be approximated well by the normal distribution. The magnitude of this change can be evaluated through simulations where the distribution $F$ used to generate the data is known.

Other computational methods for nonparametric empirical Bayes estimator are also available. Brown and Greenshtein (2009) propose a kernel method to estimate the marginal density of $\bar{v}_j$ directly when variances of $\bar{v}_j$ are all the same (for example, all teachers have the same associated sample sizes). This is based on the observation that the nonparametric Bayes estimator $\tilde{\alpha}_j$ defined in Theorem 2 does not depend directly on $F$ but rather on the marginal density of $\bar{v}_j$. When variances are homogeneous, the kernel estimator for this marginal density is easy to estimate since we have $J$ independent and identically distributed observations $(\bar{v}_1, \ldots, \bar{v}_J)$ from this marginal density.[11] Yet when individual teachers have heterogeneous variances, it is difficult to apply these methods since the observations $(\bar{v}_1, \ldots, \bar{v}_J)$ are no longer identically distributed.

## 4 Simulations

In this section, we use simulations to compare the performance of three estimators: the fixed effects estimator, the parametric linear EB estimator, and our nonparametric empirical Bayes estimator. We measure the effectiveness of these three estimators relative to an infeasible benchmark where

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

the econometrician knows the true underlying teacher quality distribution and can therefore use the optimal Bayes rule. Further, we consider the performance of these estimators under three distributions: normal, mixed normal, and chi-squared. Simulated data are based on equation (2.2) for 10,000 individual teachers with $\epsilon_{ijt} \sim \mathcal{N}(0, 0.025)$ (set to mimic what we estimate using data later).

We assess the estimators according to mean squared error. In addition, we also consider the misclassification rates of teachers in the bottom five percent in terms of both Type I and Type II error.[12] With homogeneous class sizes, every estimator is a monotone function of teacher fixed effects, guaranteeing that teacher rankings (and thus misclassification rates) are identical across the three methods. Once we allow class size to be different, however, the NPEB and EB estimators are no longer order preserving with respect to teacher fixed effects. Misclassification rates may therefore differ depending on the estimator used, although we expect that these rates should be similar so long as we consider relatively symmetric single-peaked distributions.[13] We therefore consider both a homogeneous class size case where every teacher has a total class size of 20 and a heterogeneous class size case where class size is a random draw from the set $\{20, 40\}$ with equal probability.

**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 of the EB estimator is *correct* and so it performs identically to the infeasible estimator where the distribution is known. We see that parametric EB estimator substantially improves on the mean squared error of the fixed effects estimator. Crucially, however, it only outperforms our NPEB estimator by a small margin; mean squared error is less than one percent higher under NPEB, which is quite remarkable as NPEB does not make use of any parametric assumption on the distribution $F$. As expected, misclassification rates are very similar across methods.

**Teacher Quality Distribution is Non-Normal:** Tables 1(b) and 1(c) show simulation

[12] Type I error means a teacher is ranked below 5% while her true quality ranking is above 5%. Conversely, Type II error is defined as a teacher being ranked above 5% when her true quality is below 5%.

[13] While rankings between the methodologies do not usually change substantially, they do under multi-peaked or highly skewed distributions. To see this, consider a symmetric distribution with two peaks at the first and third quartile of the distribution and a teacher with value-added located between the first quartile and the mean. On one hand, the parametric EB estimator forces the teacher towards zero (i.e., the mean), increasing her value-added. The NPEB, on the other hand, pushes the teacher towards the mass point at the first quartile, decreasing her value-added. Given that the Bayes shrinkage is going in opposite directions, teacher rankings are highly dependent on the methodology used with such a distribution.

results when teacher quality is not normally distributed. Specifically, Table 1(b) has true teacher quality following $F = 0.98\mathcal{N}(0, 0.03) + 0.01\mathcal{N}(-1, 0.03) + 0.01\mathcal{N}(1, 0.03)$ (as in equation (2.6)), while true teacher quality follows $F \sim \chi^2_1$ in Table 1(c). The mixed normal distribution has the same mean and variance as the normal distribution in Table 1(a) to create a suitable comparison.

Here, our NPEB estimator substantially outperforms *both* the parametric EB estimator and the fixed effects estimator in terms of mean squared error. Even more impressive is that our NPEB estimator achieves a mean squared error very close to that of the infeasible estimator where the true distribution is known: Our NPEB estimator has mean squared error less than one percent higher than the infeasible estimator for both distributions considered. Contrarily, the mean squared error of the parametric EB and fixed effect estimator are 15-60 percent higher than the infeasible estimator. In general, the parametric EB estimator performs worse as the underlying distribution deviates further from normality and so its efficacy is particularly poor for the chi-squared distribution. Our NPEB estimator, on the other hand, adapts to any distribution. Once again, misclassification rates are similar across all methods.

These simulations draw attention to adaptability of our nonparametric approach. Despite being a data-hungry nonparametric method, our approach performs similarly to the parametric EB even when true distribution is normal. When the true distribution is not Gaussian, our approach substantially outperforms the parametric EB estimator. Given there is no *a priori* reason to believe that teacher quality follows some specific distribution, researchers can be confident that our method adapts to any underlying distributions of teacher quality, performing almost as if the true distribution was known.

## 5 Data

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

**North Carolina:** Our first administrative data set covers all public school students in North

Carolina for fourth and fifth grades from 1996-97 to 2010-11 and third grade from 1996-97 to 2008-09.[14] These data cover roughly 1.85 million students with 4.6 million student-year observations. We also have detailed demographic data 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 grade repetition.

We follow Clotfelter, Ladd, and Vigdor (2006) and subsequent research using North Carolina data to construct the sample used to estimate teacher value-added by making several restrictions. Specifically, we require that all students are matched to a valid classroom teacher, and that students have a valid lagged test score in that subject. After the sample restrictions, our final sample consists of roughly 2.7 million student-year observations, covering 1.4 million students and 35,000 teachers.

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

**Large Urban School District:** Our second data source comes from a large urban school district in the western United States. The data span fourth and fifth grade from 2003-04 to 2012-13 and 2015-16 to 2016-17 and third grade from 2003-04 to 2012-13.[15] These data cover roughly 800,000 students with 1.7 million student-year observations. Detailed demographic data include parental education (missing for thirty percent of sample), economically disadvantaged status, ethnicity, gender, age, limited English status, and grade repetition.

Similar to the North Carolina data set, we make several restrictions to construct our value-added estimation sample. Once again, our major restrictions involve dropping students who cannot be matched to a classroom teacher, and students who do not have a valid lagged test score in the

[14] Our analysis is restricted to students in third through fifth grade since our data set 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 third grade ceases after 2008-09 because the third grade pretest was discontinued after that year.

[15] The lack of data in 2013-14 and 2014-15 is due to a change in the statewide testing regime that occurred in 2013-14, which produced 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.

relevant subject. Our value-added sample consists of 1.3 million student-year observations, covering roughly 660,000 students and 11,000 teachers.

Columns (3) and (4) of Table 2 provides summary statistics for the large urban district sample. Column (3) reports these for the entire sample, while column (4) reports them for the value-added analysis data set. Similar to the North Carolina case, we find that our value-added sample is moderately positively selected, with student test scores being about 0.06 standard deviations higher in the value-added sample.

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

# 6 Value-Added Estimates

This section reports estimates of teacher value-added from our NPEB methodology along with those using the parametric EB method. Results are described for North Carolina and the large urban school district separately. We highlight differences in the estimated teacher quality distributions among these two data sets and how these differences will in turn impact the policy analysis in Section 7. We finish by comparing the out-of-sample performance of our NPEB estimator relative to that of the parametric EB estimator.

## 6.1 North Carolina

Figure A.1(a) displays a boxplot of teacher fixed effects for teachers who appear once, twice, three times or more than three times in our North Carolina data set, respectively. Teachers showing up for more periods – typically, more experienced teachers – have less dispersion as the fixed effect for these teachers is estimated with a larger effective sample size than teachers showing up less often. The average fixed effect appears similar regardless how often teachers appear in the data (conditional on teacher experience). Bayesian shrinkage is then applied to these fixed effects, shrinking fixed effect estimates for teachers with small sample sizes back toward the mean. The boxplot in Figure A.1(c) shows the magnitude of shrinkage applied by our NPEB estimator. As expected, only small amounts of shrinkage are applied for teachers with more than three years of data, while teachers who appear less frequently – and thus have smaller effective sample sizes – are shrunk more aggressively towards zero.

The estimated teacher value-added distribution used for our NPEB (estimated nonparametrically using equation (3.1)) is shown in Figure 2(a). The estimated distribution appears at first glance to be approximately normal, although policymakers are particularly interested in the tails of the distribution. Given that, Figure 2(b) takes a cube root of the distribution in order to enhance the tails. Here, we can see that the teacher quality distribution exhibits a fat right tail relative to the Gaussian distribution. In this light, we anticipate that in North Carolina our NPEB methodology will mostly agree with the parametric EB method which assumes normality, except in the right tail.

Next, we compare our estimator with the parametric EB estimator. To do so, we start by estimating the distribution of teacher value-added under the normality assumption with maximum likelihood.[16] Using mathematics scores from all North Carolina school districts, we find that teacher value-added is distributed $\mathcal{N}(0, 0.047)$ (standard deviation = 0.217) and that $\hat{\sigma}^2_\epsilon = 0.248$. Figure 3(a) then compares the value-added estimates obtained from the parametric and nonparametric empirical Bayes estimators. In line with expectations, teacher value-added is nearly identical for teachers in the 0-95$^{th}$ percentile of the distribution, signified by the fact that nearly all dots in this region reside on the 45 degree line in Figure 3(a). Past the 95$^{th}$ percentile, however, there are some disagreements between the two methodologies, with the NPEB estimator not shrinking teachers' who have a fixed effect in the right tail as aggressively as the parametric EB estimator. Intuitively, this arises since our nonparametric methodology finds a distribution with a fatter right tail than the normal distribution assumed by the parametric method, and so it shrinks teachers in the right tail less to match the fat tail. Equivalently, this behavior can be seen in Figure A.1(a) which plots empirical Bayes estimates relative to the fixed effect estimates for a representative teacher who has

[16] This is the Chetty, Friedman, and Rockoff (2014a) no-drift estimator. Maximum likelihood is used as it is the most efficient estimator, although results are similar if we use method of moments instead.

taught twenty students: the NPEB and parametric EB estimates sit on top of each other in the left tail (until the extreme left tail of the distribution is reached), but deviate substantially in the right tail, with the NPEB estimates being substantially above those estimated using the parametric EB methodology.

In terms of policy, the right tail of the distribution is likely to considerably impact the expected policy gains from interventions targeting high value-added teachers. So while much of the teacher quality distribution appears to be normally distributed in North Carolina, the misspecification of the normality assumption in the right tail may lead to widely different policy evaluations between the parametric and nonparametric methodologies.

### 6.2 Large Urban School District

Figures A.1(a) and A.1(c) present boxplots of teacher fixed effects and the magnitude of shrinkage that our NPEB estimator applies by how often teachers appear in the data. These figures are similar to those for North Carolina, although the large urban district appears to feature larger variance in teacher fixed effects, which leads to the district featuring higher dispersion.

Our estimated teacher value-added distribution using NPEB is shown in Figure 2(c). Notably, the distribution features higher dispersion and greater skewness than in the North Carolina data. Enhancing the tails in Figure 2(d) yields other compelling differences: the value-added distribution has a truncated left tail. In contrast, North Carolina featured a left tail similar to one from the Gaussian distribution. Our NPEB methodology should therefore mostly agree with the parametric EB method, except in the left tail.

Using parametric EB, we find teacher value-added is distributed $\mathcal{N}(0, 0.0977)$ (standard deviation $= 0.3126$) and that $\hat{\sigma}_{\epsilon}^2 = 0.2596$, which is considerably higher than the variance we found in North Carolina. Comparing the value-added estimates obtained from the two empirical Bayes estimators, Figure 3 indicates that teacher value-added is very similar for teachers in the 5-$100^{th}$ percentile of the distribution. For teachers in the left tail, however, our nonparametric method shrinks them far more aggressively back toward the mean since our nonparametrically estimated distribution features a truncated left tail. Correspondingly, the representative teacher plotted in Figure A.1(b) has similar NPEB and parametric EB estimates in the right tail, but has NPEB estimates above the parametric EB estimates in the left tail beginning at the $5^{th}$ percentile of the distribution, which is important for policy predictions as there is still substantial mass at this point (relative to the extreme tails).

These differences in the tails of the value-added distribution relative to a normal distribution can drive large differences in policy calculations between the two empirical Bayes methodologies. On one hand, the normality assumption is particularly misspecified in the right tail for North Carolina. Given that this is the set of teachers impacted from interventions targeting high value-added teachers, we expect that the gains from these policies, such as retention bonuses, will be *understated* in the parametric EB methodology since it shrinks high value-added teachers back towards the mean too forcefully. On the other hand, the normality assumption is misspecified in the left tail for the large urban district, which is crucial for policies targeting low value-added teachers, such as teacher releases. Here, the parametric EB methodology is likely to *overstate* the benefit of these policies since it does not sufficiently shrink these low value-added teachers back toward the mean.

### 6.3 Out-of-Sample Prediction of Teacher Performance

Given our estimates of value-added above, we now evaluate the performance of our NPEB estimator relative to the parametric EB and fixed effects estimators. A natural evaluation of the estimators is their ability to predict future outcomes. For instance, imagine that school boards or policymakers observe a teacher's past performance and want to predict future outcomes for that teacher. We measure the performance of this prediction via the squared error distance: $(y_{j,t+1} - \hat{y}_{j,t+1})^2$, where $y_{j,t+1}$ is the true outcome of teacher $j$ in period $t+1$ and $\hat{y}_{j,t+1}$ its predictor, utilizing all past information relating to her teaching performance, starting when the teacher first appears in the sample up until the $t$-th period.

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

(TMSE) and normalized mean squared error (NMSE). TMSE is given by:

$$TMSE = \frac{1}{N_j} \sum_{j \in I_j} \left( (y_{j,t+1} - \hat{y}_{j,t+1})^2 - \frac{\sigma_\epsilon^2}{n_{j,t+1}} \right), \tag{6.1}$$

where $I_j$ is the set of teachers whose performance is being predicted and $N_j$ is the size of the set. Without the adjustment term $\frac{\sigma_\epsilon^2}{n_{j,t+1}}$, the quantity is nothing but the usual sum of squared errors; the adjustment term is introduced to account for the effect of different class sizes on the variance. Alternatively, we use NMSE:

$$NMSE = \frac{1}{N_j} \sum_{j \in I_j} \left( n_{j,t+1} (y_{j,t+1} - \hat{y}_{j,t+1})^2 \right), \tag{6.2}$$

which is the usual sum of squared errors with an adjustment term $n_{j,t+1}$. The adjustment term scales back the contribution of teachers with large classes sizes – who have better precision by construction – by the size of their class, $n_{j,t+1}$, to resolve the fact value-added is easier to predict among these teachers.

Tables 3(a) and 3(b) report TMSE and NMSE using the North Carolina and large urban district data for three different estimators: NPEB, parametric EB, and fixed effects. Each row in the table represents the number of years of past information relating to teacher $j$'s performance used for the prediction.[17] The empirical Bayes methods substantially outperform the fixed effects method when only a few years of prior data are used. As more and more years of data become available, the gain from using empirical Bayes diminishes in comparison with the fixed effect estimator. The extra gain is quite substantial, however, when information about a teacher is scarce.

Comparing the two empirical Bayes estimators, the NPEB dominates the parametric EB under both prediction accuracy measures, except when only using one prior year of information with the large school district data set. When 3-5 years of data are used, the prediction performance of the NPEB estimator surpasses that of the parametric EB by the largest margin. The NPEB estimator is best with a few years of data as it needs sufficient data for its superiority to become apparent. With many years of data, the superiority of the NPEB then declines since it no longer shrinks the

[17] To predict the performance of teacher $j$ using $t$ years of data, we restrict the sample to include teachers who appear $t$+1 times. For example, when using three periods of data to predict the performance of a teacher, we subset the sample to include only teachers appearing at least four times.

fixed effects estimates materially and so it approaches the performance of the fixed effect estimator. Importantly, our nonparametric methodology outperforms that of the parametric EB by the highest margin with 3-5 years of data per teacher, exactly when teacher tenure decisions are often made by districts.

# 7 Implications for Policy

This section performs policy calculations for policies targeting the bottom and top of the teacher quality distribution. We pay particular attention to differences in policy calculations found using our nonparametric method relative to those found using the parametric EB methodology.

## 7.1 Policy Experiment I: Lay-off Policies

One policy recommendation that has gained considerable traction is to replace poor-quality teachers with mean-quality teachers. The specific proposal made by Hanushek (2009, 2011) and further explored by Chetty, Friedman, and Rockoff (2014b) involves replacing the bottom 5% of the teachers with those who are of average quality. Here, we calculate the policy gains from replacing the bottom q% of teachers, although we pay particular attention to the bottom 5% policy given the considerable scrutiny it has received in the prior literature.

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

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

where $\Phi$ is the CDF of $\mathcal{N}(0, \sigma_\alpha^2)$ and $q$ is the cutoff percentage. Similarly, the total gain in test scores for a policy that deselects the bottom q% of teachers, denoted $TR(q)$, is:

$$TR \equiv \mathbb{E}[\alpha \mathbb{1}\{\alpha \geq \Phi^{-1}(q)\}]\,, \tag{7.2}$$

where $\mathbb{1}\{\alpha \geq \Phi^{-1}(q)\}$ takes the value one if the teacher's quality is greater than the cutoff value $\Phi^{-1}(q)$ and zero otherwise.

Under the assumption that $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, it is easy to derive that

$$MR(q) = -\frac{\int_{-\infty}^{\Phi^{-1}(q)} \alpha\phi(\alpha)d\alpha}{\int_{-\infty}^{\Phi^{-1}(q)} \phi(\alpha)d\alpha} \tag{7.3}$$

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

with $\phi(\alpha)$ being the PDF of $\mathcal{N}(0, \sigma_\alpha^2)$.

The normality assumption underpinning the parametric EB estimator, however, is likely misspecified given the distributions we nonparametrically estimated in Section 6. Under our nonparametrically estimated distribution, replacement teachers remain mean zero (i.e., $\alpha = 0$) since value-added is always centered. With our estimated distribution, however, the marginal and total test score gains from a policy laying off the bottom $q$ percent of teachers is instead:

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

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

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

Theorem 1 states that the distribution $F$ is identified from the data and so we can evaluate the policy under two distributions of value-added: (i) the normal distribution with a standard deviation estimated from the data, and (ii) the distribution $F$ estimated using the same data set.

Figure 4 compares the policy gains under the parametric EB methodology where it is assumed that $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$ to those found using our NPEB method where we let $\alpha \sim F$. For North Carolina, the parametric EB methodology overestimates policy gains, although not substantially.

The large urban district paints a different picture, however, with the parametric EB methodology considerably overestimating the policy gains.

Tables 4(a) and 4(b) report test score gains under the parametric EB methodology and our NPEB methodology for a policy that releases the bottom $q\%$ of teachers in North Carolina and the large urban district, respectively. Focusing on the bolded policy to release bottom vigintile teachers, columns (1) and (2) indicate that the parametric EB method overstates the policy gains by about five percent relative to our NPEB method. The normality assumption thus appears well-founded in North Carolina when considering policies targeting the left tail. Against this, the parametric EB method overstates the policy gains by *twenty-four* percent in the large urban district. Here, the normality assumption is clearly misplaced, highlighting the importance of using our flexible method that adapts to the underlying distribution, since this distribution is unknown *a priori*.

**Accounting for Fact Value-Added is Estimated:** The above policy analysis was based on the fact that we know true teacher value-added and thus can accurately distinguish teachers based on their quality. In reality, teacher value-added is estimated. To account for this, we replace the test score gains in equations (7.3) and (7.5) with their sample analogs.[18] These sample analogs are calculated via Monte Carlo simulation 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. For instance, under the parametric EB methodology we first sample 40000 observations from $\mathcal{N}(0, \sigma_\alpha^2)$. Second, for each sample observation we generate the noisy data $y_j = \alpha_j + \epsilon_j$ where $\epsilon_j \sim \mathcal{N}(0, \sigma_\epsilon^2/(k \cdot n))$, where $n$ represents yearly class sizes (set at 20) and $k$ represents the number of years of data for each teacher (set at 3). Third, we use the parametric EB estimator to obtain an estimated VA, $\hat{\alpha}_j$, and calculate $\frac{1}{40000}\sum_j \alpha_j 1\{\hat{\alpha}_j > \hat{\Phi}_{\hat{\alpha}}^{-1}(q)\}$ as an estimator for $\hat{TR}$ (sample analog of equation (7.4)). By the law of large numbers, this produces a consistent estimator for $\hat{TR}$. Analogously, we also sample value-added from distribution $F$, use the nonparametric empirical Bayes method to obtain an estimator $\hat{\alpha}_j$, and calculate $\hat{TR}$ in a similar fashion.

Results based on these simulations are presented in Figure A.2. The policy gains fall when using

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

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

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

estimated rather than true value-added since some teachers with true value-added below the fifth percentile are kept. The decrease in test score gain, however, is relatively subdued and is similar for both the parametric EB and NPEB methodologies. This is unsurprising since the methodologies do not substantially affect the ranking of teachers and so using estimated rather than true value-added should impact them in a similar manner.

Columns (3) and (4) of Tables 4(a) and 4(b) report these policy gains when true teacher value-added is unobserved. Results are very similar to the case when value-added was observed to the policymaker: the parametric EB method overstates the policy gains by eight percent in North Carolina and by *twenty-seven* percent in the large urban district.

### 7.2 Policy Experiment II: Retention Policies

We can also consider designing policy that focuses on the teachers that have quality in the upper quantile of the distribution. The though experiment is then how much marginal and total test score gain we would have by retaining these top quality teachers in the school district versus losing them. For the comparison, we maintain the assumption that if the high quality teachers are not retained, we would need to replace them by an average quality teacher (with zero value added). Suppose the policy is to be designed to retain teachers that have quality greater and equal to the $1 - q$ percentile of the quality distribution, we can define the gains as

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

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

Computation details are analogous to Policy I and the results using North Carolina maths test data and the large urdan district are reported in Figure **??** - **??**.

## 8 Conclusion

In this paper, we have proposed a new approach to estimating teacher value-added that relaxes the normality assumption featuring in the popular parametric Empirical Bayes method. Our non-parametric Empirical Bayes approach is appealing in that it allows us to estimate the underlying distribution of teacher quality directly, and is computationally feasible using large data sets.

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

The general nonparametric approach to estimation has broader applicability to other areas of research in education, where the underlying heterogeneity of students, teachers and schools is intrinsic. For example, looking beyond the current application, our methodology is well-suited to capture dynamic policy-driven changes in underlying teacher quality distributions. Suppose, for instance, policymakers implemented a policy releasing teachers in the bottom of the teacher value-added distribution *every* year. Under such a policy, the left tail of the teacher quality distribution would become truncated. Imposing the normality assumption, 'fitting the data' would then require lowering the value-added of teachers near the truncation point to create a left tail. The value-added of teachers at the bottom of the distribution would therefore be underestimated, in turn implying that the gains of continuing the policy could be substantially overestimated. Given that our method estimates such changes in the underlying teacher quality distributions flexibly, it provides policymakers with sharper predictions as to the extent of continuing policy gains when implementing such dynamic reforms.

# References

Abadie, A., and M. Kasy (forthcoming): "Choosing among regularized estimators in empirical economics: The risk of machine learning," *Review of Economics and Statistics*.

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

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

——— (2014b): "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," *American Economic Review*, 104(9), 2633–79.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2006): "Teacher-student matching and the assessment of teacher effectiveness," *Journal of Human Resources*, 41(4), 778–820.

Delaigle, A., and A. Meister (2008): "Density estimation with heteroscedastic error," *Bernoulli*, 14(2), 562–579.

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

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

Evdokimov, K., and H. White (2012): "Some extensions of a Lemma of Kotlarski," *Econometric Theory*, 28, 925–932.

Gu, J., and R. Koenker (2017a): "Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data," *Journal of Applied Econometrics*, 32(3), 575–599.

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

Gu, J., and S. Shen (2017): "Oracle and adaptive false discovery rate controlling methods for one-sided testing: Theory and application in treatment effect evaluation," *Econometrics Journal*, 21(1), 11–35.

Hanushek, E. A. (2009): "Teacher Deselection," in *Creating a New Teaching Profession*, ed. by D. Goldhaber, and J. Hannaway, pp. 165–180. Urban Institute Press, Washington, DC.

——— (2011): "The economic value of higher teacher quality," *Economics of Education Review*, 30(3), 466–479.

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

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

Jacob, B. A., and L. Lefgren (2008): "Can principals identify effective teachers? Evidence on subjective performance evaluation in education," *Journal of Labor Economics*, 26(1), 101–136.

Jiang, W., and C.-H. Zhang (2009): "General maximum likelihood empirical Bayes estimation of normal means," *Annals of Statistics*, 37(4), 1647–1684.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008): "What does certification tell us about teacher effectiveness? Evidence from New York City," *Economics of Education Review*, 27(6), 615–631.

Kane, T. J., and D. O. Staiger (2008): "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," Working Paper 14607, National Bureau of Economic Research.

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

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

Koenker, R., and I. Mizera (2014): "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules," *Journal of the American Statistical Association*, 109(506), 674–685.

Kotlarski, I. (1967): "On Characterizing the Gamma and the Normal Distribution," *Pacific Journal of Mathematics*, 20, 69 – 76.

Laird, N. (1978): "Nonparametric maximum likelihood estimation of a mixing distribution," *Journal of the American Statistical Association*, 73(364), 805–811.

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

Morris, C. N. (1983): "Parametric empirical Bayes inference: Theory and applications," *Journal of the American Statistical Association*, 78(381), 47–55.

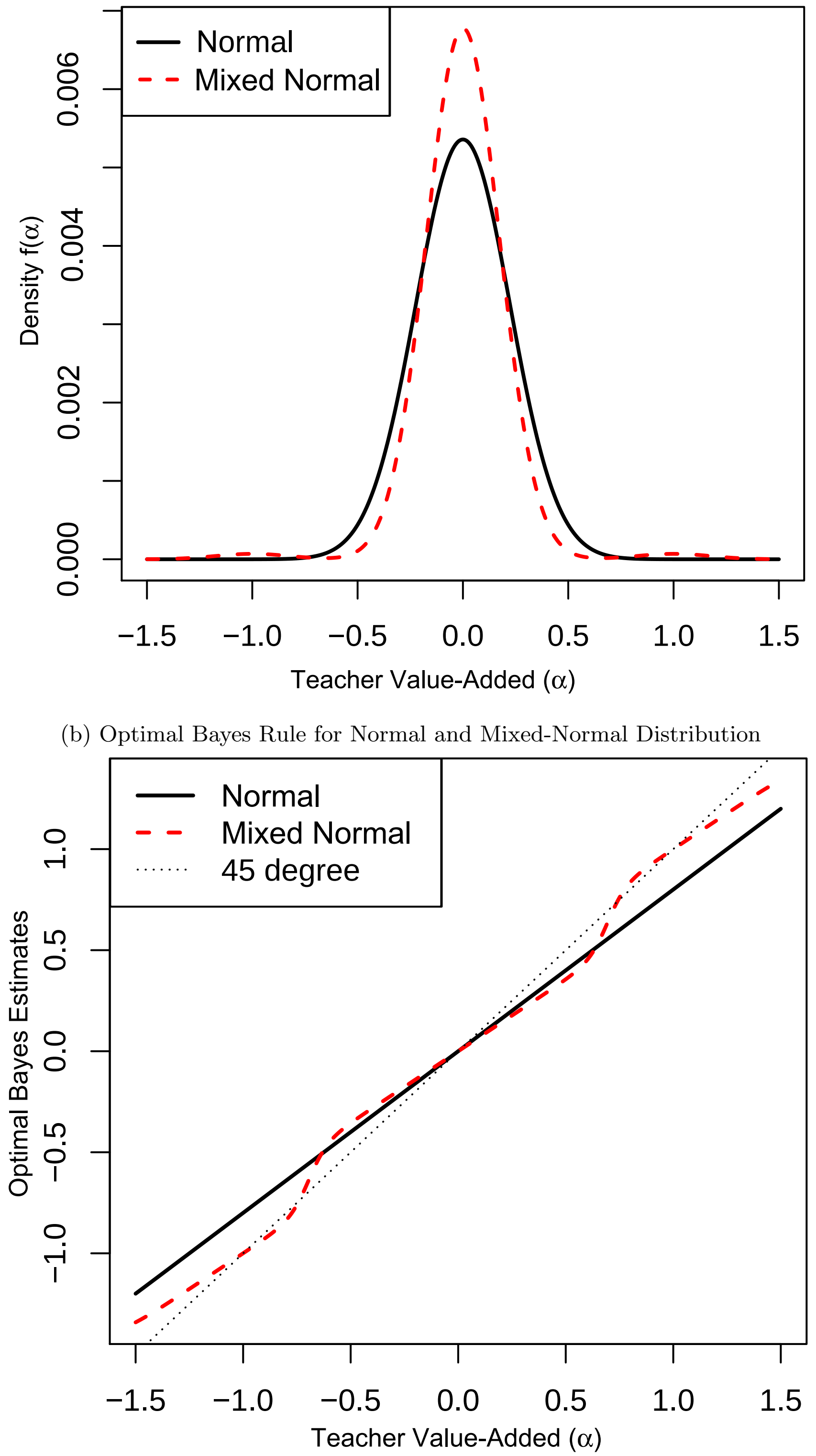
Rao, B. (1992): *Identifiability in Stochastic Models: Charaterization of Probability Distributions.* Academic Press, United Kindom.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005): "Teachers, schools, and academic achievement," *Econometrica*, 73(2), 417–458.

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

Rockoff, J. E. (2004): "The impact of individual teachers on student achievement: Evidence from panel data," *American Economic Review*, 94(2), 247–252.

Figure 1: Example: Bayes Estimators for Normal and Mixed-Normal Distributions

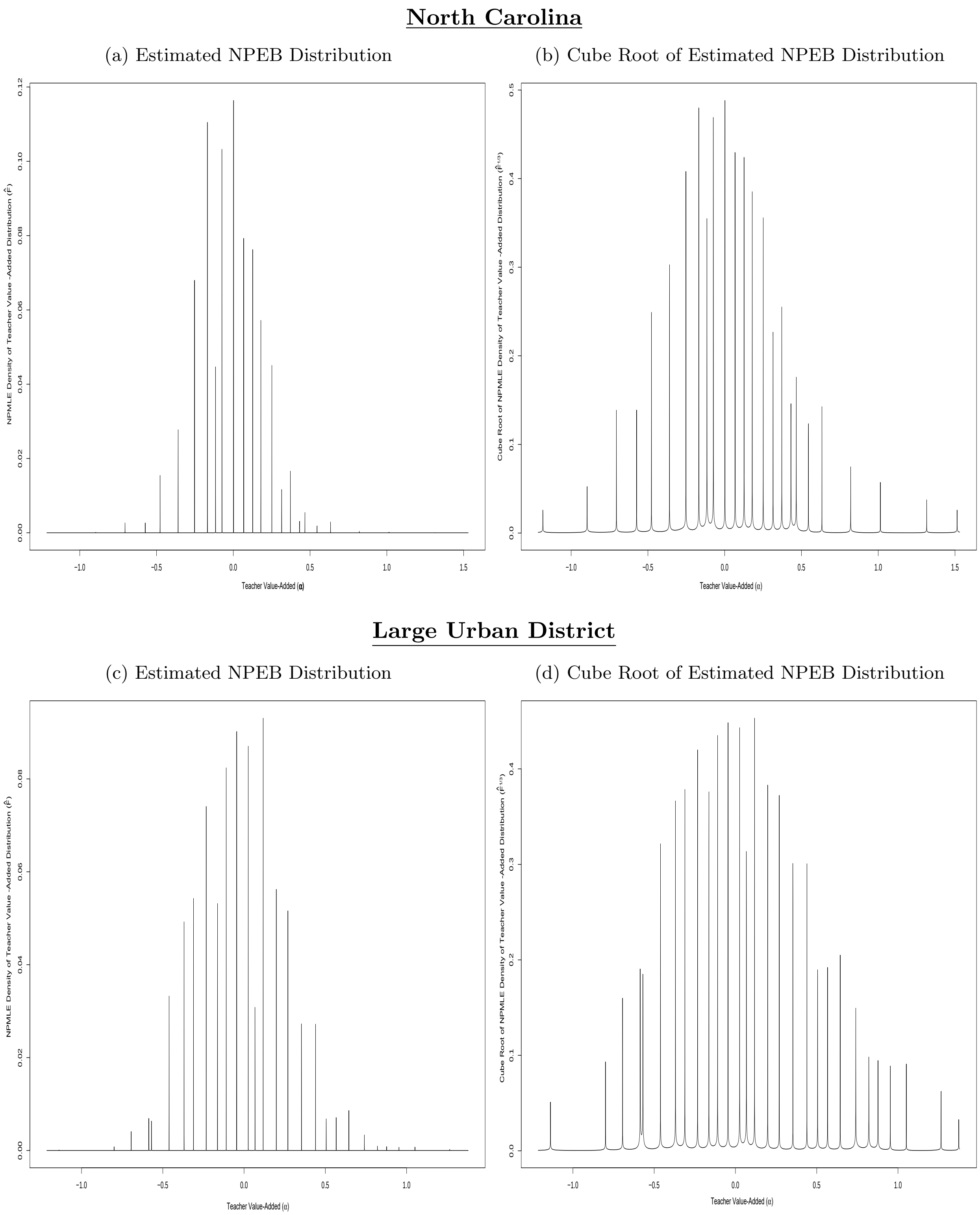


(a) PDFs of Normal and Mixed Normal Distributions

(b) Optimal Bayes Rule for Normal and Mixed-Normal Distribution

Notes: Figure 1(a) plots the density function of the two types of distribution in our example in Section 2: the solid black line is the normal distribution with mean zero and variance 0.05 ($\alpha_j \sim \mathcal{N}(0, 0.05)$), and the dashed red line corresponds to the mixed normal distribution specified in equation (2.6), given by $\alpha_j \sim 0.98\mathcal{N}(0, 0.03) + 0.01\mathcal{N}(-1, 0.03) + 0.01\mathcal{N}(1, 0.03)$. Figure 1(b) then shows the optimal Bayes estimates for these two distributions, with the black solid line representing the normal distribution and the dashed red line depicting the mixed normal distribution. The dotted line is the 45 degree line which indicates where the optimal Bayes is the same under the two underlying distributions.

Figure 2: Estimated Teacher Quality Distributions Using Nonparametric Empirical Bayes



**North Carolina**

(a) Estimated NPEB Distribution

(b) Cube Root of Estimated NPEB Distribution

**Large Urban District**

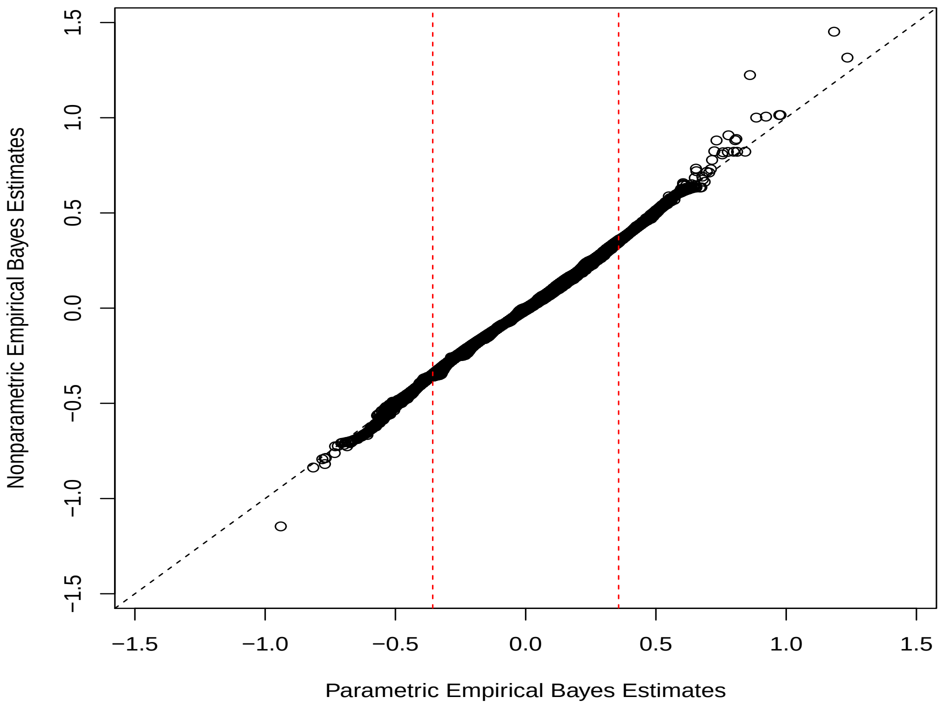
(c) Estimated NPEB Distribution

(d) Cube Root of Estimated NPEB Distribution

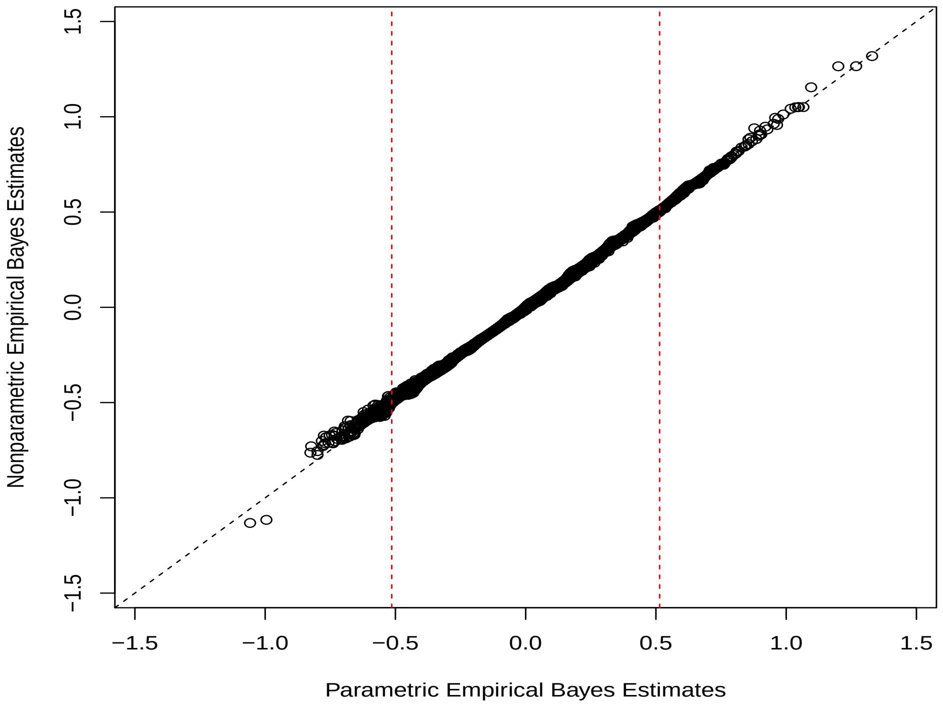
Notes: Figures 2(a) and 2(c) display the estimated distribution of teacher quality, $\hat{F}$, for North Carolina and the Large Urban District, respectively. These distributions are estimated nonparametrically using equation (3.1). In order to better see tail behavior, Figures 2(b) and 2(d) take a cube root of the estimated value-added distribution to boost the tails of the distribution for North Carolina and the Large Urban District, respectively.

Figure 3: Comparison of Value-Added Estimates Using Parametric and Nonparametric Empirical Bayes

(a) North Carolina

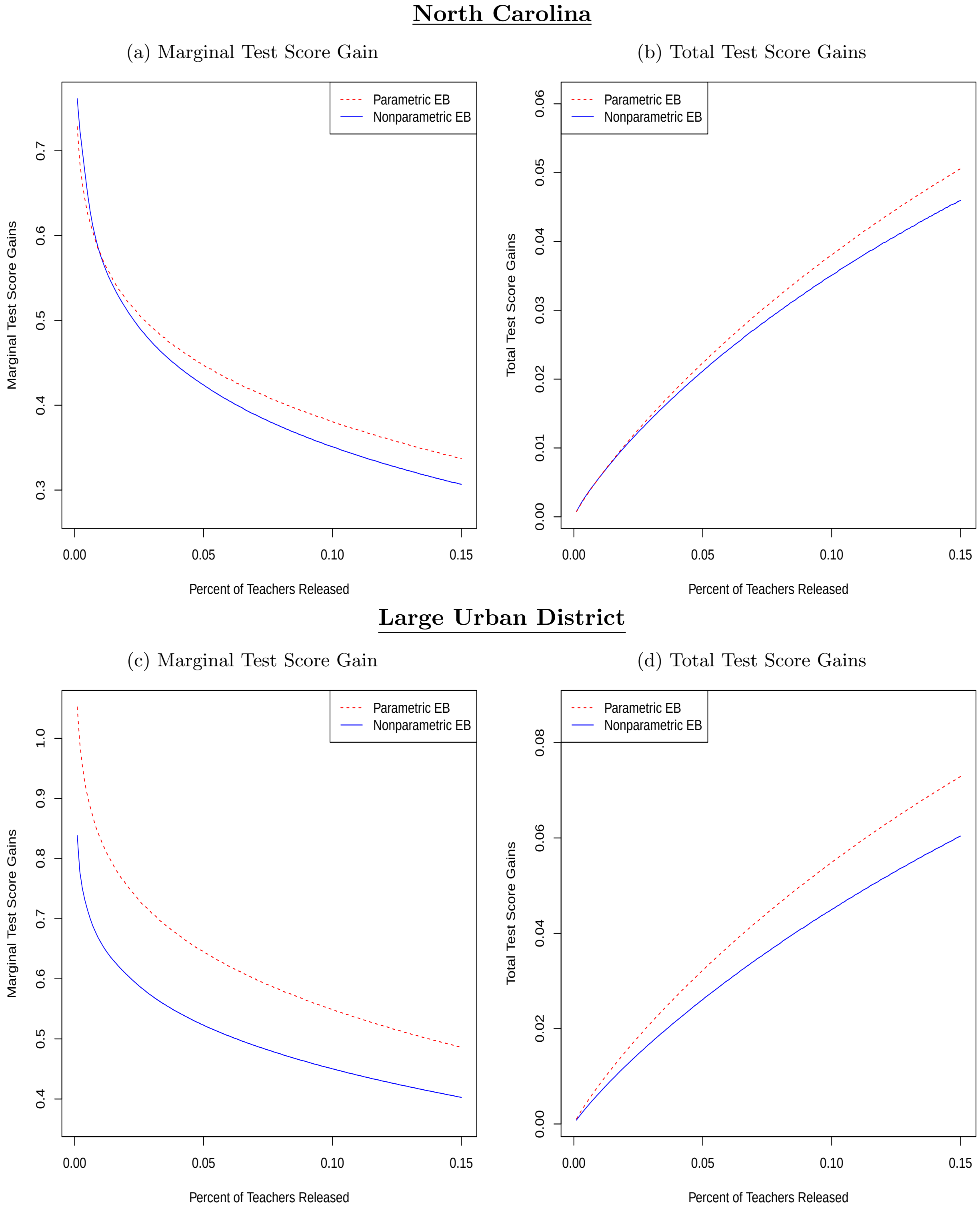


(b) Large Urban District



Notes: Figures 3(a) and 3(b) compare estimates of teacher value-added obtained from the parametric EB estimator (on the x-axis) and our NPEB estimator (on the y-axis) for North Carolina and the Large Urban District, respectively. The 45 degree line is represented with a dashed line and indicates where the two estimators agree in terms of teacher value-added estimates. The vertical dashed lines represent the $5^{th}$ and $95^{th}$ percentiles of teacher value-added estimates according to the parametric EB estimator to delineate the tails of the value-added estimates.

Figure 4: Test Scores Gains from Replacing Bottom $q$ Percentile of Teachers



**North Carolina**

(a) Marginal Test Score Gain

(b) Total Test Score Gains

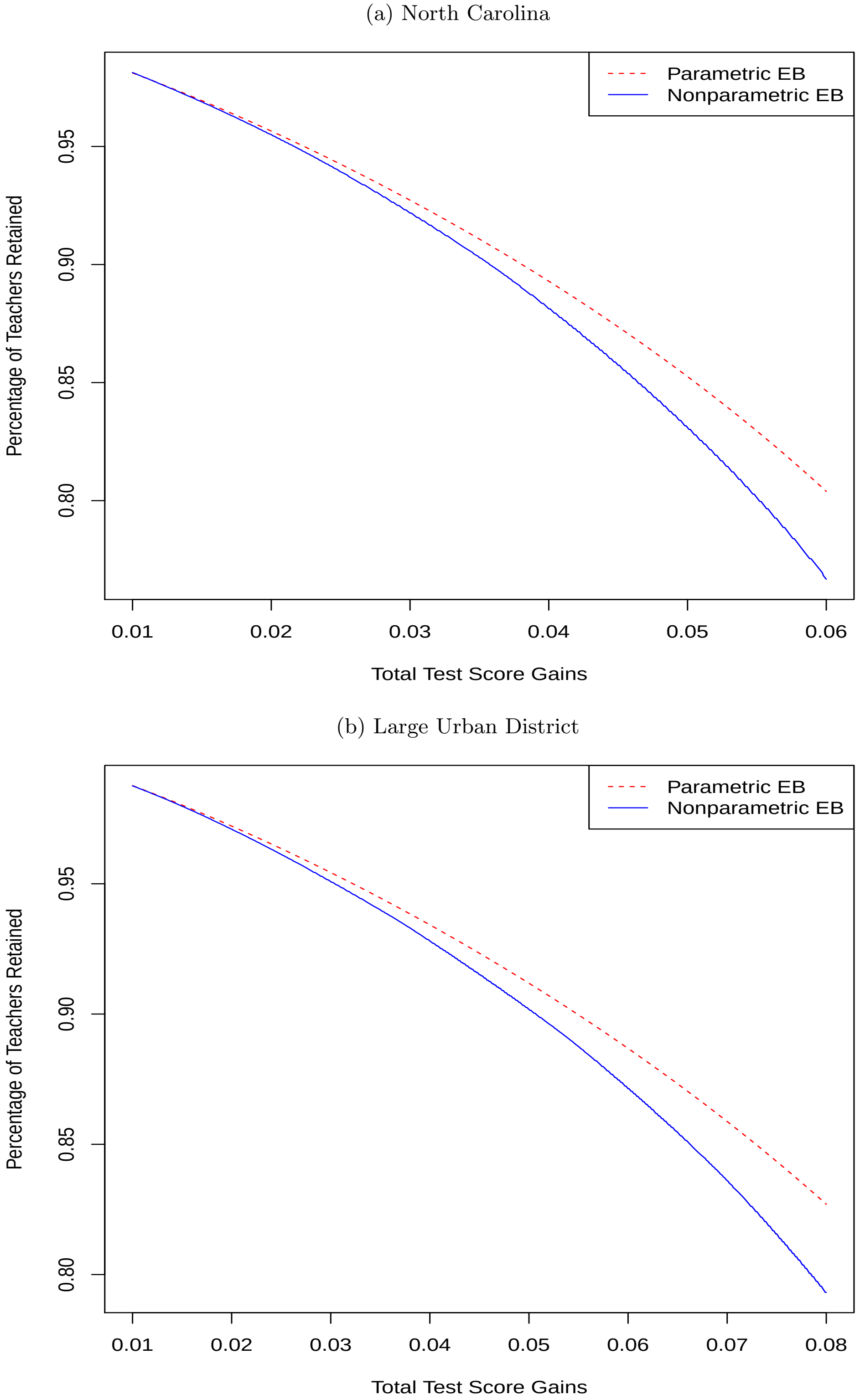
**Large Urban District**

(c) Marginal Test Score Gain

(d) Total Test Score Gains

Notes: Figures 4(a) and 4(b) show the marginal and total test score gains of a policy that releases the bottom $q\%$ of teachers in North Carolina, while figures 4(c) and 4(d) do the same for the large urban district. The dotted lines indicate the policy gains expected under the parametric EB methodology where it is assumed that $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, where $\sigma_\alpha^2$ is estimated from the data. The solid lines denote the policy gains expected under our NPEB methodology where we allow $\alpha \sim F$, with $F$ being estimated directly from the data using equation (3.1). Policy gains reported here assume that policymakers observe true underlying value-added, although Figure A.2 presents the estimated gains if value-added is estimated rather than observed.

Figure 5: Test Scores Gains from Retaining top $q$ Percentile Teachers


(a) North Carolina

(b) Large Urban District

Notes: Figures 5(a) and 5(b) display the total test score gains from retaining teachers above the $q^{th}$ percentile of the value-added distribution in North Carolina and the large urban district, respectively. The dashed lines indicate the policy gains expected under the parametric EB methodology where it is assumed that $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, where $\sigma_\alpha^2$ is estimated from the data. The solid lines denote the policy gains expected under our NPEB methodology where we allow $\alpha \sim F$, with $F$ being estimated directly from the data using equation (3.1). Policy gains reported here assume that policymakers observe true underlying value-added, although Figure A.3 presents the estimated gains if value-added is estimated rather than observed.

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

| | Homogeneous Class Sizes (Total Class Size of 20) | | | | Heterogeneous Class Sizes (Total Class Size 20-40) | | | |
|---|---|---|---|---|---|---|---|---|
| | Infeasible | NPEB | EB | FE | Infeasible | NPEB | EB | FE |
| *Mean Squared Error* | 10.81 | 10.87 | 10.81 | 12.50 | 11.03 | 11.07 | 11.04 | 12.84 |
| Teacher ranked in bottom (top) 5% when true VA above (below) 5% (Type I error)<br>(Type I error=Type II error by definition) | | | | | | | | |
| *Bottom 5%* | 152.5 | 152.5 | 152.5 | 152.5 | 153.5 | 153.6 | 153.5 | 153.9 |
| *Top 5%* | 152.6 | 152.6 | 152.6 | 152.6 | 153.6 | 153.5 | 153.6 | 154.1 |

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

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

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

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

Notes: Tables 1(a), 1(b) and 1(c) use simulation to compare the performance of four estimators when the distribution of teacher value-added follows a normal, mixed normal and a chi-squared distribution, respectively. Specifically, Table 1(a) has teacher value-added being normally distributed with mean zero and variance 0.05 (i.e., $F \sim \mathcal{N}(0, 0.08)$), while Table 1(b) has true teacher quality following $F \sim 0.98\mathcal{N}(0, 0.03) + 0.01\mathcal{N}(-1, 0.03) + 0.01\mathcal{N}(1, 0.03)$ (as in the example given by equation (2.6)). The normal and mixed normal distributions have the same mean and variance to create a suitable comparison. Teacher value-added follows $F \sim \chi^2_1$ in Table 1(c). The infeasible estimator is the optimal estimator given that the true distribution is known to the econometrician (which is infeasible as it is unknown in practice), the nonparametric emiprical Bayes (NPEB) estimator which nonparametrically estimates the underlying distribution, the parametric empirical Bayes (EB) estimator which assumes that the underlying distribution is normal, and the fixed effect (FE) estimator which applies no empirical Bayes shrinkage. The simulation averages results from 500 repetitions with 10,000 individual teachers setting $\sigma^2_\epsilon = 0.025$. Results are reported for homogeneous class sizes where every teacher has a class size of twenty students and heterogeneous class sizes where class sizes of teachers are a random draw from the set $\{20, 40\}$ with equal probability. Note that teacher rankings are identical for the three methods under homogeneous class sizes. Only Type I error (teacher ranked in bottom (top) 5% when true VA above (below) 5%) is reported as it is identical to that of Type II error (teacher ranked above (below) 5% when true VA below (above) 5%).

Table 2: Summary Statistics

| | North Carolina | | Large Urban District | |
|---|---|---|---|---|
| | Full Sample[1] | Value-Added Sample | Full Sample[2] | Value-Added Sample |
| | (1) | (2) | (3) | (4) |
| *Mean of Student Characteristics* | | | | |
| Math Score ($\sigma$) | 0.00 | 0.05 | 0.00 | 0.07 |
| Reading Score ($\sigma$) | 0.00 | 0.03 | 0.00 | 0.06 |
| Lagged Math Score ($\sigma$) | 0.01 | 0.03 | 0.03 | 0.08 |
| Lagged Reading Score ($\sigma$) | 0.01 | 0.03 | 0.03 | 0.07 |
| % White | 57.8 | 60.1 | 9.3 | 9.1 |
| % Black | 28.8 | 27.9 | 9.9 | 8.6 |
| % Hispanic | 7.4 | 6.5 | 74.0 | 75.5 |
| % Asian | 2.0 | 1.9 | 4.3 | 4.4 |
| % Free or Reduced Price Lunch[3] | 46.3 | 44.6 | 77.9 | 78.2 |
| % English Learners | 4.3 | 3.5 | 28.0 | 28.9 |
| % Repeating Grade | 1.5 | 1.5 | 1.5 | 0.4 |
| Parental Education:[4] | | | | |
| % High School Dropout | 11.5 | 10.6 | 34.5 | 34.4 |
| % High School Graduate | 47.31 | 47.0 | 27.6 | 27.8 |
| % College Graduate | 25.4 | 25.9 | 20.1 | 20.0 |
| Teacher Experience:[5] | | | | |
| 0-2 Years of Experience | 18.6 | 18.8 | 4.9 | 4.8 |
| 3-5 Years of Experience | 15.3 | 15.6 | 10.5 | 10.3 |
| # of Students | 1,847,615 | 1,386,555 | 810,753 | 664,044 |
| # of Teachers | 76,503 | 35,053 | 15,267 | 11,078 |
| Observations[6] (student-year) | 4,562,218 | 2,680,027 | 1,707,459 | 1,280,569 |

[1] North Carolina data coverage: grades 4-5 from 1996-97 through 2010-11 and grade 3 from 1996-97 through 2009-10.

[2] Large Urban District data coverage: grades 4-5 from 2003-04 through 2012-13 and 2015-16 through 2016-17 school years and third grade from 2003-04 through 2012-13.

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

[4] Omitted category is some college and college graduate also incorporates those with graduate school degrees. For North Carolina this data is missing after the 2005-06 school year, while thirty percent of observations in the large urban district are missing parental education data or have parental education recorded as "Decline to Answer".

[5] Omitted category is greater than five years of experience. For the full sample, teacher experience data is missing for about twenty and fifteen percent of observations for North Carolina and the large urban school district, respectively.

[6] Data is 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 are missing for about one percent of observations. For the large urban district (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 (Parametric EB) and Fixed Effects (North Carolina Data)

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

Table 3(b): Predicted Performance of Nonparametric Empirical Bayes (NPEB), Parametric Empirical Bayes (Parametric EB) and Fixed Effects (Large Urban School District Data)

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

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

Table 4(a): Test Scores Gains of Policy Releasing Bottom $q$% Teachers (North Carolina Data)

| | True Value-Added Observed | | True Value-Added Unobserved | |
|---|---|---|---|---|
| % Teachers Released | Test Score Gain under $F$ | Test Score Gain under Normal | Test Score Gain under $F$ | Test Score Gain under Normal |
| | (NPEB) | (EB) | (NPEB) | (EB) |
| | (1) | (2) | (3) | (4) |
| 1 | 0.0058 | 0.0058 | 0.0056 | 0.0059 |
| 2 | 0.0103 | 0.0105 | 0.0099 | 0.0104 |
| 3 | 0.0142 | 0.0147 | 0.0137 | 0.0145 |
| 4 | 0.0178 | 0.0187 | 0.0171 | 0.0183 |
| **5** | **0.0212** | **0.0224** | **0.0203** | **0.0219** |
| 6 | 0.0243 | 0.0258 | 0.0233 | 0.0252 |
| 7 | 0.0272 | 0.0291 | 0.0261 | 0.0284 |
| 8 | 0.0300 | 0.0322 | 0.0288 | 0.0314 |
| 9 | 0.0326 | 0.0352 | 0.0312 | 0.0342 |
| 10 | 0.0351 | 0.0381 | 0.0336 | 0.0369 |

Table 4(b): Test Scores Gains of Policy Releasing Bottom $q$% Teachers (Large Urban School District Data)

| | True Value-Added Observed | | True Value-Added Unobserved | |
|---|---|---|---|---|
| % Teachers Released | Test Score Gain under $F$ | Test Score Gain under Normal | Test Score Gain under $F$ | Test Score Gain under Normal |
| | (NPEB) | (EB) | (NPEB) | (EB) |
| | (1) | (2) | (3) | (4) |
| 1 | 0.0066 | 0.0083 | 0.0064 | 0.0086 |
| 2 | 0.0121 | 0.0151 | 0.0118 | 0.0153 |
| 3 | 0.0171 | 0.0213 | 0.0166 | 0.0213 |
| 4 | 0.0217 | 0.0269 | 0.0211 | 0.0269 |
| **5** | **0.0261** | **0.0323** | **0.0254** | **0.0322** |
| 6 | 0.0302 | 0.0372 | 0.0294 | 0.0371 |
| 7 | 0.0342 | 0.0420 | 0.0332 | 0.0417 |
| 8 | 0.0379 | 0.0465 | 0.0369 | 0.0461 |
| 9 | 0.0415 | 0.0507 | 0.0405 | 0.0503 |
| 10 | 0.0450 | 0.0549 | 0.0438 | 0.0543 |

Notes: Tables 4(a) and 4(b) display the estimated gains in mathematics scores in terms of student level standard deviations of a policy that releases the bottom $q$% of teachers and replaces them with mean quality teachers for North Carolina and the large urban district, respectively. 'Test score gain under $F$' reports the test score gain of the policy when teacher quality is distributed according the distribution $F$ – nonparametrically estimated using equation (3.1) – and applying the NPEB estimator to calculate value-added. 'Test score gain under normal' reports the test score gain when teacher quality is normally distributed and the parametric EB estimator is used to calculate teacher value-added. The left panel assumes that true teacher value-added is observed by the policymaker and is the same as the policy gain shown in Figure 5. Since true value-added is unknown to the policymaker, the right panel represents the policy gains when value-added is estimated. These gains are the same as those Figure A.3 and are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. The bolded line indicates the widely-analyzed release bottom five percent teachers policy.

# A General Deconvolution with Panel Data

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

## A.1 Teacher value-added 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 $f(\cdot, \cdot)$, $\alpha$ is a random variable with pdf $g(\cdot)$, and $\epsilon_1$ and $\epsilon_2$ are random variables from the same pdf $h(\cdot)$ with mean zero.*

**Assumption 2** *$\alpha$, $\epsilon_1$, and $\epsilon_2$ are mutually independent.*

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

**Lemma 1 (Kotlarski (1967))** *Under Assumption 1-3, the pdf's of $\alpha$ and $\epsilon$ 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 $\alpha$ and $\phi_\epsilon(t)$ the characteristic function of $\epsilon$, 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

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

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

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

where the second equality holds because $\epsilon_1$ 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 $\epsilon_1$ and $\epsilon_2$ 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

$$\hat{\psi}(u,v) = \frac{1}{J}\sum_{j=1}^{J}\exp(iuy_{1j} + ivy_{2j})$$

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

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

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

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

## A.2 Teacher value-added 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 $f(\cdot,\cdot,,\cdot,\cdot)$, $\alpha$ is a random variable with pdf $g(\cdot)$, $\theta_1$ and $\theta_2$ are random variables from the same pdf $q(\cdot)$ with mean zero and $\epsilon_{11}$, $\epsilon_{12}$, $\epsilon_{21}$, and $\epsilon_{22}$ are random variables from the same pdf $h(\cdot)$ with mean zero.*

**Assumption 5** *$\alpha$, $\theta_1$, $\theta_2$, $\epsilon_{11}$, $\epsilon_{12}$, $\epsilon_{21}$, and $\epsilon_{22}$ are mutually independent.*

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

**Lemma 2** *Under Assumptions 4 - 6, the pdf's of $\alpha$, $\theta$ and $\epsilon$ 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 $\epsilon$ and the joint distribution $(Y_{12},Y_{22})$ uniquely determines the pdf of $Z_2$ and $\epsilon$. Now letting the characteristic function of $(Y_{11},Y_{12})$ be denoted as $\psi_{Y_{11}Y_{12}}(t_1,t_2)$, we have

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

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 $\phi_\epsilon$, the characteristic function of $(Z_1, Z_2)$ is therefore identified. Now apply Lemma 1 again on

$$Z_1 = \alpha + \theta_1$$

$$Z_2 = \alpha + \theta_2$$

to identify the densities of $\alpha$ and $\theta$. $\blacksquare$

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 $\alpha$, $\theta_{jt}$, and $\epsilon_{ijt}$ are mutually independent of each other, the pdf's of $\alpha, \theta$, and $\epsilon$ are nonparametrically identified.

# B Construction of the Teacher Value-Added Sample

This appendix describes the construction of the final sample of students and teachers used for teacher value-added estimation in both of our administrative datasets. Our sample selection follows that of prior work (for instance, Chetty, Friedman, and Rockoff (2014a,b)) and the main requirements to be in the sample is 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 of that subject.

## B.1 North Carolina

For North Carolina, we follow Clotfelter, Ladd, and Vigdor (2006) and subsequent research using North Carolina data to construct our sample. We start with the entire enrollment history of students in North Carolina for grades 4-5 over the 1996-97 through 2010-11 school years and grade 3 over the 1996-97 through 2009-10 school years.[19] These data cover roughly 1.85 million students with 4.6 million student-year observations.

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

We then make several sample restrictions. First, we drop the 1.37 million student-year observations who we have identified as having an invalid teacher. This is by far our biggest sample restriction and the majority of the sample restriction comes from the fact we assign teachers to students based on who proctors the student's exam. To ensure the teacher proctoring the exam is the same as the classroom teacher, we confirm that the proctor is teaching a primary grade math and English class. If the teacher is not, we drop the observations as we are no longer confident that we are correctly matching classes to teachers. Second, we drop charter school classrooms and

[19] 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.

special education classrooms leading to a loss of an addtional 70,000 student-year observations. Third, we drop 16,000 observations where we lack data on teacher experience. Fourth, we exclude 380,000 observations that lack a valid current or lagged test score in that subject, with half of this loss coming from a lack of third grade math pretest data in 2005-06 and third grade English pretest data in 2007-08 due to a statewide test update.[20] Fifth, we only include classes with more than seven but fewer than forty students with valid current and lagged test scores in that subject, creating a loss of 10,000 observations.[21] Our final sample is roughly 2.7 million student-year observations, covering 1.4 million students and 35,000 teachers.

### B.2 Large Urban School District

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

For demographics, we have information on 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 approximately twenty-nine percent of the sample. Whenever parental education is missing, we create a missing indicator for that demographic variable.

We then make several sample restrictions. First, we drop 100,000 student-year observations who we cannot match to a teacher. Second, we drop 1800,000 observations where we lack data on teacher experience. The data we drop here is over-represented in early years since we only have teacher experience data from 2007-08 onward.[22] 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

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

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

[22] We infer teacher experience data from 2003-04 through 2006-07 from the observed teacher experience 2007-08, but we cannot get teacher experience teacher 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 and about 3-5% every year thereafter.

score in that subject.[23] Our final sample is roughly 1.3 million student-year observations, covering roughly 660,000 million students and 11,000 teachers.[24]

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

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

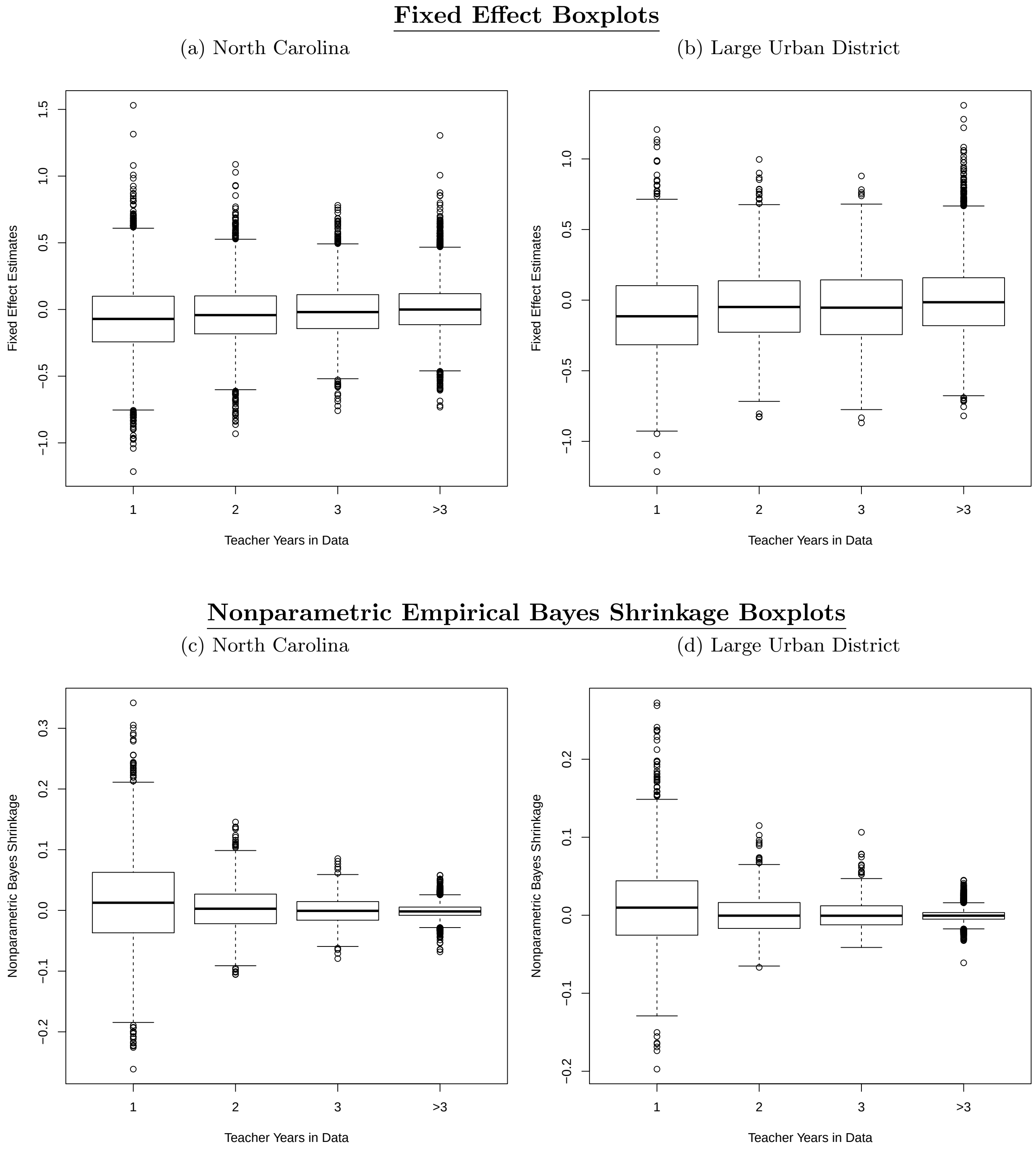
We follow Chetty, Friedman, and Rockoff (2014a,b) and parametrize the control function for lagged test scores $f_{1g}(A_{i,t-1})$ with a cubic polynomial in prior-year scores in math and English and interact these cubics with the students's grade level. When prior test scores in the other subject are missing, we set the other subject prior score to zero and include an indicator for missing data in the other subject interacted with the controls for prior own-subject test scores. We parametrize the control function for teacher experience $f_2(e_{j(i,g,t)})$ using dummies for years of experience from 0 to 5, with the omitted group being teachers with 6 or more years of experience. The student-level control vector $X_{igt}$ consists of the respective demographic variables in each dataset. The class-level control vector $\bar{X}_{c(i,g,t)}$ includes class size, cubics in class and school-grade means of prior-year test scores in math and English each interacted with grade, class and school-year means of all the individual covariates $X_{igt}$, and (4) grade and year dummies.

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

[24] Data on teacher experience is only available in this large urban school district from 2007-08 onward. We can include teacher experience at a loss of 160,000 student-year observations.

# APPENDIX FIGURES AND TABLES

Figure A.1: Boxplots of Fixed Effects and Nonparametric Empirical Bayes Shrinkage



Fixed Effect Boxplots

(a) North Carolina

(b) Large Urban District

Nonparametric Empirical Bayes Shrinkage Boxplots

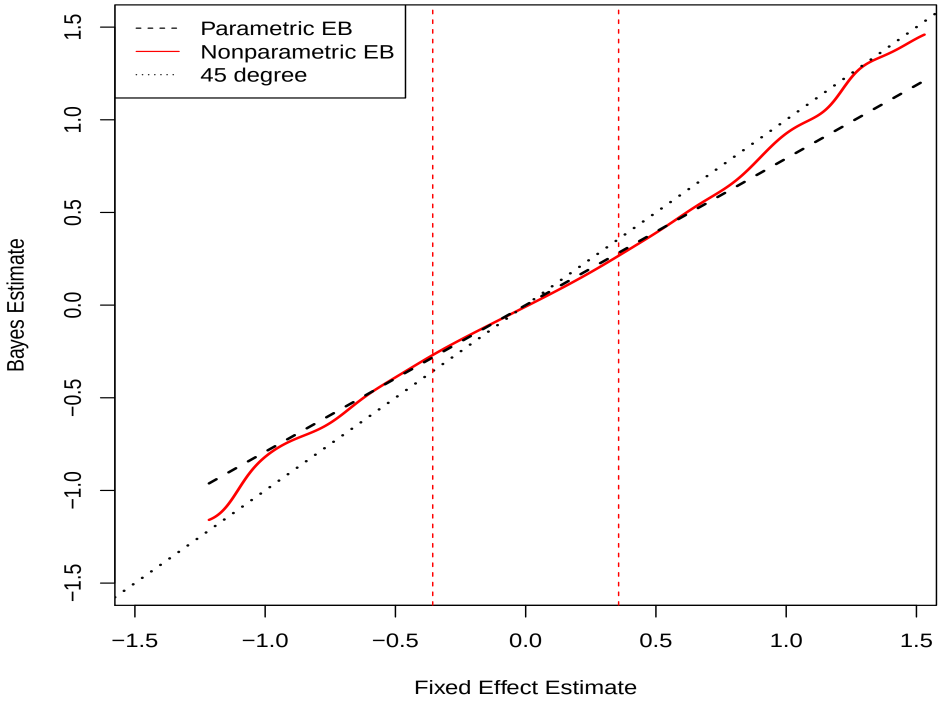
(c) North Carolina

(d) Large Urban District

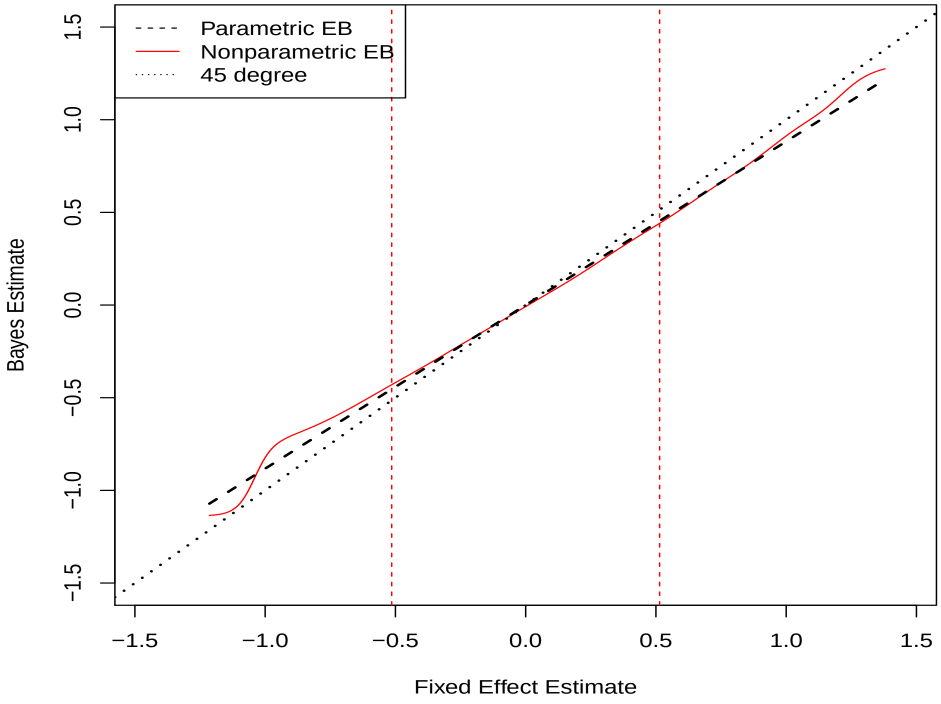
Notes: These figure give the raw fixed effect estimates and the value of shrinkage applied using our nonparametric empirical Bayes methodology by the number of times a teacher appears in the data. Specifically, Figures A.1(a) and A.1(b) display a boxplot of fixed effect estimates for teachers who appear once, twice, three times or more than three times in our North Carolina and Large Urban District datasets, respectively. Similarly, Figures A.1(c) and A.1(d) show a boxplot of the absolute value of shrinkage applied to teachers who appear once, twice, three times or more than three times in our North Carolina and Large Urban District datasets, respectively. Boxplots use the box to indicate the interquartile range between the first and third quartile and use whiskers to indicate the first (third) quartile minus (plus) the interquartile range multiplied by 1.5. Outliers outside this range are illustrated with dots.

Figure A.1: Empirical Bayes 'Shrinkage' for Fixed Total Class Size (total class size = 20)

(a) North Carolina

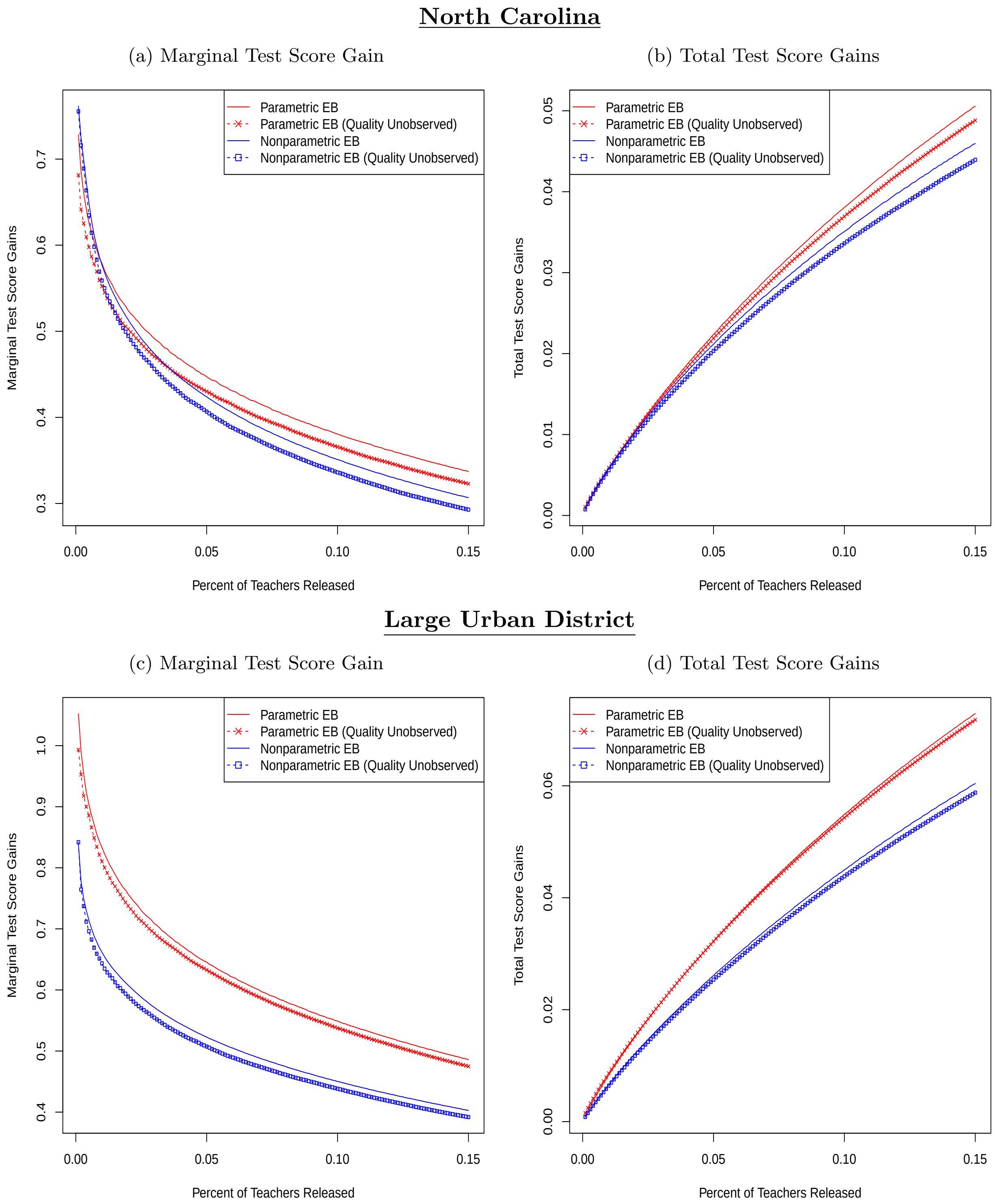


(b) Large Urban District



Notes: Figures A.1(a) and A.1(b) show fixed effect estimates relative to the Bayes estimates for both the nonparametric empirical Bayes (NPEB) and the parametric empirical Bayes (parametric EB) methodologies. The dotted line represents the 45 degree line and indicates where the fixed effect and empirical Bayes estimates agree. Since the amount of Bayes 'shrinkage' applied depends on the total number of students taught by the teacher, we display the rule for a representative teacher who has taught a total class size of twenty students throughout her career. The vertical dashed lines represent the $5^{th}$ and $95^{th}$ percentiles of teacher value-added estimates according to the fixed effect estimates to delineate the tails of the value-added estimates.

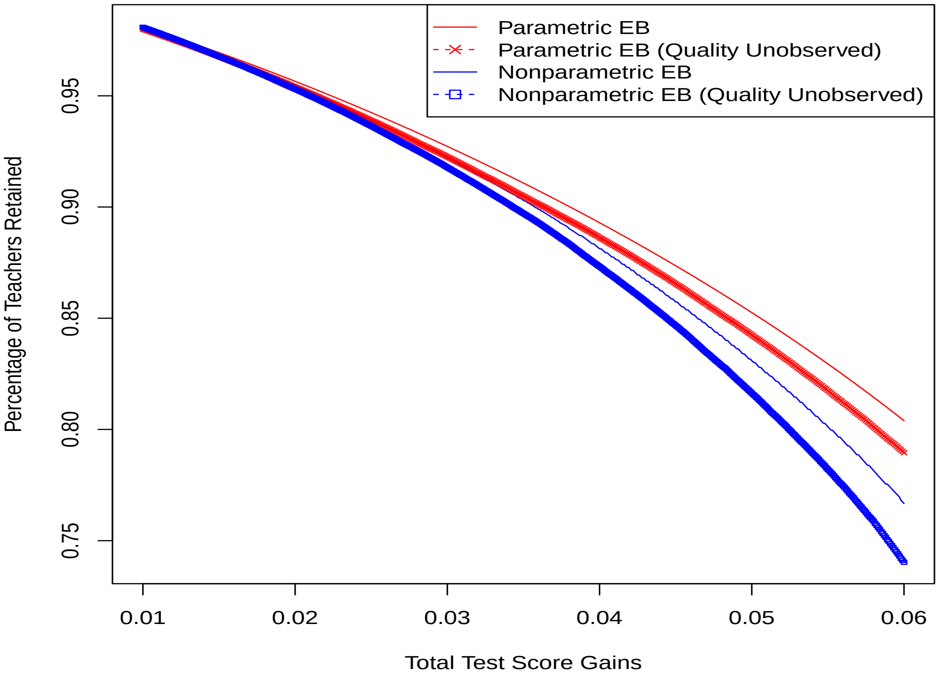
Figure A.2: Test Scores Gains from Replacing Bottom $q$ Percentile of Teachers when Value-Added is Estimated



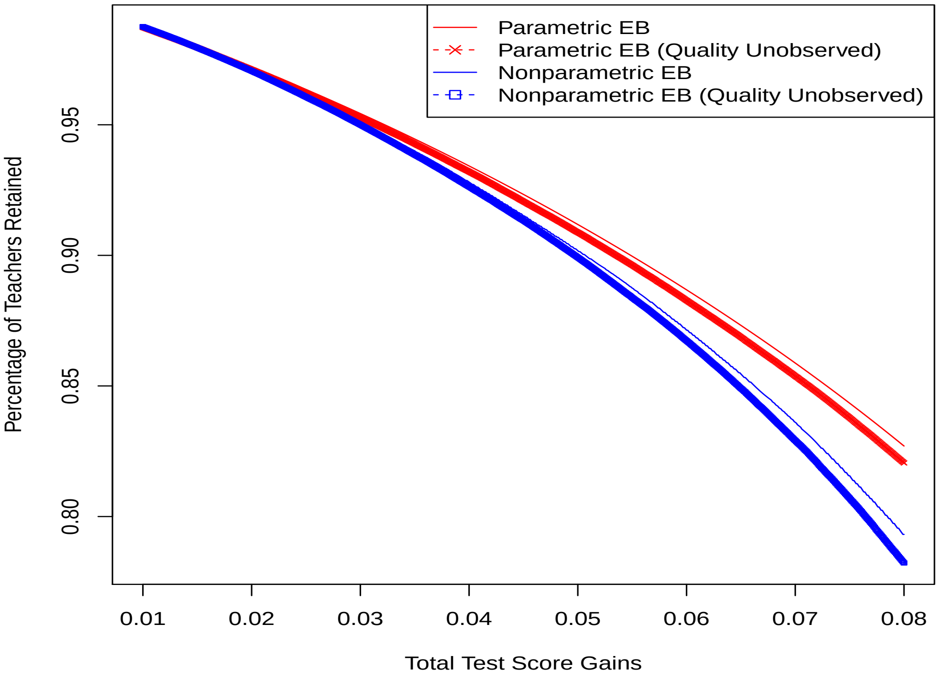
Notes: Figures A.2(a) and A.2(b) show the marginal and total test score gains of a policy that releases the bottom $q$% of teachers in North Carolina, while figures A.2(c) and A.2(d) do the same for the large urban district. The solid lines indicate the policy gains expected under the parametric EB and NPEB methodology when true teacher value-added is observed and are identical to those presented in Figure 4. The dashed lines represent the policy gains when value-added is estimated. These gains are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. Details of the simulation are provided in Section 7.1.

Figure A.3: Test Scores Gains from Retaining top $q$ Percentile Teachers when Value-Added is Estimated

(a) North Carolina



(b) Large Urban District



Notes: Figures A.3(a) and A.3(b) display the total test score gains from retaining teachers above the $q^{th}$ percentile of the value-added distribution in North Carolina and the large urban district, respectively. The solid lines indicate the policy gains expected under the parametric EB and NPEB methodology when true teacher value-added is observed and are identical to those presented in Figure 5. The dashed lines represent the policy gains when value-added is estimated. These gains are calculated via Monte Carlo simulation with 40,000 observations under the assumption that we estimate $\hat{\Phi}_{\hat{\alpha}}$ and $\hat{F}_{\hat{\alpha}}$ using three years of data for each teacher and assuming teachers all have class sizes of twenty. Details of the simulation are provided in Section 7.2.