2014

2014-3 Targeting the Wrong Teachers: Estimating Teacher Quality for Use in Accountability Regimes

Nirav Mehta

Targeting the Wrong Teachers: Estimating Teacher Quality for Use in Accountability Regimes

by

Nirav Mehta

Working Paper # 2014-3 **July 2014**

Western

CIBC Working Paper Series

Department of Economics
Social Science Centre
Western University
London, Ontario, N6A 5C2
Canada

This working paper is available as a downloadable pdf file on our website
http://economics.uwo.ca/centres/cibc/

Targeting the Wrong Teachers: Estimating Teacher Quality for Use in Accountability Regimes

Nirav Mehta
University of Western Ontario *

July 10, 2014

Abstract

This paper compares the performance of popular estimators of teacher quality, which serve as inputs into teacher incentive schemes. I model an administrator tasked with categorizing teachers with respect to an exogenous cutoff, showing that the preferred estimator depends on the relationship between teacher quality and class size. I then use data from Los Angeles to show that the simpler fixed effects estimator would outperform the more popular empirical Bayes estimator, meaning that the administrator would prefer to use it to either reward high-performing teachers or sanction low-performing ones. The preferred estimator would create 200 fewer classification errors.

*nirav.mehta@uwo.ca I thank Tim Conley, Steven Glazerman, Lance Lochner, Rachel Margolis, Henry May, and Todd Stinebrickner for useful discussions pertaining to this paper, and the SSHRC Insight Development Grant program for funding.

The design and introduction of incentive pay schemes for teachers is a linchpin of education policy reform. The vast majority of teacher remuneration is based on their experience and credentials (Podgursky and Springer (2006, 2011)), but the fact that only a small amount of variation in student achievement is explained by experience and credentials (Hanushek (1986), Goldhaber and Brewer (1997), Rivkin et al. (2005)) has spurred a debate towards introducing incentive pay for teachers. Recently, this debate has moved from theory into actual education policies affecting large numbers of students and teachers, such as President Obama's Race to the Top initiative (McGuinn (2012)), or the TAP program, which has introduced performance-based bonuses to over 20,000 teachers serving over 200,000 students across the US.[1]

Correctly providing incentives in an environment where teachers may vary in both inherent effectiveness and unobserved effort is a difficult contracting problem, which has caused existing schemes to adopt several simplifications. First, student test score gains are assumed to separately depend on teacher quality and other inputs, resulting in a value-added (VA) model (Hanushek (1979)). Second, high-stakes schemes often take the form of cutoff rules that reward (punish) teachers with estimated VA above (below) some cutoff. For example, Glazerman et al. (2011) document that about half of the performance-based schemes drawing on the Teacher Improvement Fund are based on cutoff rules.[2] Finally, to create tractable schemes that can be implemented using standard statistical software, decisions are typically based on either fixed effects (FE) estimates of teacher VA or posterior means of teacher quality ("empirical Bayes" estimates (EB)). The EB estimator is a weighted average of the FE estimator and the average quality amongst teachers being compared under the incentive scheme (such as those in a school district or state), and is derived by first positing a population distribution of teacher quality, which serves as a prior belief about teacher quality, and then updating this prior with student achievement data using Bayes' rule. In one extreme, when no observations are associated with a particular teacher the posterior mean is the same as the population mean (i.e. the mean of the prior distribution of teacher quality). On the other hand, when an arbitrarily large amount of data is available for a teacher, the EB estimator converges to the FE estimator and the prior receives no weight.[3]

Inferences about teacher quality that are based on finite samples are inherently noisy. Practitioners often find the noisiness of such estimators unattractive, especially when they are used to make high-stakes decisions like retention, promotion, or bonuses, because there may be a non-negligible probability of making an error based on an estimator of teacher quality. For example, American Federation of Teachers President Randi Weingarten said in a 2012 interview about releasing VA scores to the public: "I fought against it because we knew value-added was based on a series of assumptions and not ready for prime-time. But back then, we didn't realize the error rates could be as high as 50 percent!" (Goldstein (2012)). EB estimators are favored by many practitioners and education researchers because they minimize the mean squared error; that is, they reduce the variance in estimated teacher quality by putting positive weight on the more precisely estimated mean of teacher VA for the relevant set of teachers, at the cost of introducing statistical bias. McCaffrey et al. (2003) write "Early [value-added model] applications (for instance, Murnane, 1975, and Hanushek, 1972) primarily used fixed effects, while more recent applications (including the TVAAS layered model) have used random effects almost exclusively" (McCaffrey et al. (2003), page 64). Economics researchers (e.g. Rockoff (2004), Kane et al. (2008)) also routinely use EB estimators in studies of teacher quality. However, there has been little guidance to date about

[1] `http://www.tapsystem.org/`

[2] More generally, Stiglitz (1991) and Ferrall and Shearer (1999) note that real world incentive schemes are often simple in shape.

[3] The EB estimator is equivalent to a "random effects" model of teacher VA.

which estimator policymakers should use.[4]

This paper compares the performance of FE and EB estimators, within the context of how they are typically used in education policy. Remarkably, most existing comparisons of these estimators adopt a statistical, not economic approach (see Section 1). Therefore, I first formalize an objective function for an administrator, such as an education policymaker or school district superintendent, tasked with identifying teachers above or below a certain level in the teacher quality distribution.[5] This approximates the discrete nature of many pay-for-performance schemes as well as more austere policies, such as firing teachers below some quality threshold. I assume the administrator is risk-neutral and chooses a cutoff policy for each estimator to minimize a weighted sum of expected Type I and Type II errors. Intuitively, her expected utility is equal to the expected probability of correct classifications, making it natural to compare the estimators based on their expected number of mistakes.[6] Because FE and EB estimators differ only by how much they weigh teacher fixed effects, where the weights depend on how many students are assigned to teachers, I compare the performance of FE and EB estimators, allowing the administrator to choose a cutoff policy to maximize her expected objective. Endogenous cutoff policies are important because it seems reasonable to allow the administrator to choose a threshold based on the distribution of estimates.[7]

This paper contains theoretical and quantitative findings that are relevant for administrators, education policymakers, and researchers wishing to correctly categorize teachers with respect to a threshold. The theoretical finding is that the relative performance of the FE and EB estimators depends on both the desired cutoff and the relationship between teacher quality and class size. The estimators perform equally well when class size is independent of teacher quality, because the EB estimator shrinks all teachers towards the population mean by the same proportion, which preserves the rankings of teachers. However, if principals would like to shift students away from the lowest quality teachers or assign the highest quality teachers to teach small classes of gifted students then class size may depend on teacher quality (e.g. Lazear (2001)), causing the relative performance of FE and EB estimators to diverge. To see why, consider a comparison between two teachers who are both of below-average quality. If higher quality teachers are assigned more students, EB estimators will put more weight on their students' test score gains and less weight on the population mean of teacher quality, relative to low-quality teachers assigned fewer students. In the extreme scenario where all but one student in a large school are assigned to a high quality teacher, both estimators for that teacher converge to the true value while the EB estimator for a low-quality teacher will likely be close to the population mean of teacher quality. Therefore, the EB estimator is likely to determine that the low-quality teacher is better. I show that the performance of FE and EB estimators differs most at the tails of the distribution of teacher quality under several plausible scenarios. This divergence is important if we seek to identify either very low- or high-performing teachers, even in the case where only a small number of teachers reside in the

[4]It is important to note that the estimators considered here may already have conditioned on variables such as class size; the key difference between the FE and EB estimators is not on how to calculate the sample mean of a teacher's student achievement growth (FE) that feeds into the estimate, but rather on how much to weigh the sample mean versus population information about teacher quality.

[5]I refer to "quality" and VA interchangeably.

[6]Risk-neutrality is a useful benchmark for several additional reasons. First, it may be a priori reasonable to assume risk-neutrality for an administrator classifying an extremely large number of teachers because there would not be a large degree of sampling variation in the administrator's objective. Second, I do not know of any work estimating the degree of risk-aversion for an administrator tasked with a similar problem in the same context, so there is no guidance as to what to use. Third, seminal theoretical work (e.g. Lazear and Rosen (1981) and Green and Stokey (1983)) assumes that the principal is risk-neutral. Finally, though the EB estimator would be preferred by an administrator with a quadratic loss function, there is no reason to believe this appropriately characterizes her degree of risk aversion, even were her loss function continuous.

[7]As I show later, in some cases this may allow her to obtain the same objective value under both estimators.

tails.[8]

I then use the developed model to quantify the prospective performance of the estimators in the Los Angeles Unified School District, the second-largest school district in the US, and a district with a large degree of diversity and variation in both student achievement and class size. Because the model highlights the importance of the relationship between teacher quality and class size, first I make another contribution by documenting this relationship, using teacher VA estimates provided by the LA Times (Buddin (2011)). I then solve the model to examine the relative performance of FE and EB estimators using values calibrated from Schochet and Chiang (2012) and the relationship I document between teacher quality and class size. I do this by solving for optimal policy cutoffs under both FE and EB estimators over the entire population of teachers, to approximate the objective of an administrator considering implementing a district-wide incentive pay scheme for a wide range of desired cutoffs. Student achievement for the average teacher in LA is about equal parts signal and noise, making it difficult for the administrator to categorize teachers with respect to an exogenous cutoff. The key finding is that the FE estimator performs better than the EB estimator for almost every desired cutoff. That is, the FE estimator would be preferred by an administrator wishing to minimize mistakes in rewarding high-quality teachers with bonuses or sanctioning low-quality teachers. In particular, I find that the administrator would make at least 200 fewer classification errors by switching from the empirical Bayes to the fixed effects estimator to categorize teachers as being in the bottom or top 1% in Reading in the Los Angeles Unified School District. Though this may represent a small fraction of teachers in Los Angeles, the recent public outcry about a case where VA scores was incorrectly calculated for 40 teachers in Washington DC, which resulted in at least one firing (Strauss (2013)), suggests that the public would also be concerned about misclassifying an even larger number of teachers in LA.

Motivated by the quantitative results showing the choice of estimator is an important component of an incentive scheme, I review existing incentive schemes and find that it is often difficult to discern the methodology used. Appendix A summarizes 15 existing teacher incentive schemes that are based in part on VA. Most of the schemes use cutoff rules to assign bonuses. More than half of the schemes base bonuses on VA models of student achievement, and almost 90% of those use EB estimators in their calculation of teacher VA. Strikingly, about one fifth of the schemes do not even specify how student achievement is mapped into teacher bonuses. A corollary of the quantitative result is that, because the choice of estimator matters, teacher incentive pay programs should indicate exactly how student achievement enters bonus assignment or personnel decisions.

1 Related Literature

VA models are the workhorse of existing teacher incentive schemes and education research. Due to their pervasiveness, I take the use of VA models as given, and examine how the most commonly used estimators of VA perform when the underlying technology is consistent with a VA model. Therefore, the focus of this paper differs from that of the wide body of research studying how effectively VA models measure teacher quality. Baker and Barton (2010), Guarino et al. (2012a), Glazerman et al. (2010), and McCaffrey et al. (2003) highlight general problems with VA estimation and the use of VA estimates in policy. Value-added models are a restricted form of a more general production technology for cognitive achievement (Todd and Wolpin (2003)), and many authors have tested these restrictions to determine whether they are good measures of teacher quality, with mixed results. First, some authors have compared estimates of teacher VA with and without random assignment of students to teachers (Kane and Staiger (2008)) or with subjective ratings of teacher effectiveness (Jacob and Lefgren (2005)), surmising that VA models do a reasonably good

[8] In 2010, the former Washington D.C. Schools Chancellor Michelle Rhee fired 241 teachers based on performance measures (Turque (2010)). Hanushek (2011) quantifies how replacing teachers at the bottom of the estimated quality distribution would affect student achievement and, ultimately, economic output.

job of measuring teacher quality. Second, there is also concern that VA models do not condition on sufficiently rich information about household or other inputs (Rothstein (2009, 2010), Andrabi et al. (2011)), though evidence is mixed here as well (Kinsler (2012), Chetty et al. (2013)). Bond and Lang (2013) question whether VA should be ascribed any cardinal meaning at all, noting that monotonic transformations of test scores can eliminate growth in the black-white reading test score gap. This paper does not contribute to this literature. Rather, because existing schemes make such extensive use of them, I take as given the underlying assumptions of the VA framework and study the much less understood question of which estimator to use *within* the VA framework.

Although many studies test the statistical validity of VA models, none compare how different estimators of VA perform from the perspective of a utility-maximizing administrator. This paper contributes to this literature by showing when different estimators would be preferred by an administrator, and computing the preferred estimator using an existing dataset for a large US school district. Comparing these estimators is crucial, given the dominant role VA models and these estimators play in policy. This paper is most closely related to Schochet and Chiang (2012), which calculates error rates for FE and EB estimators of teacher quality, assuming the same cutoff policy for both estimators. There are two salient contributions of this paper, relative to Schochet and Chiang (2012). First, I embed the analysis within an economic decision problem, as opposed to the hypothesis-testing framework used in Schochet and Chiang (2012), and Second, I allow class sizes to depend on teacher quality, which as demonstrated later can drive a wedge between the performance of the FE and EB estimators. Tate (2004) notes that ranks formed by FE and EB estimates may differ, depending on class size, but does not embed the analysis within a decision problem. In contrast, this paper offers guidance about which estimator would be preferred by a utility-maximizing decision maker, and computes the preferred estimator using an existing dataset. In a paper with a different focus, Guarino et al. (2012b) compare the performance of FE and EB estimators, with a focus on how they perform when students are not randomly assigned to teachers. Guarino et al. (2012a) compare the performance of FE and other estimators of VA, but do not consider EB estimators because they assume all teachers have the same class sizes (implying that EB estimates are FE estimates times a constant). This paper's contributions relative to Guarino et al. (2012a) is first showing in Section 2.1 that the estimators perform equally well when all teachers have the same class size, and then demonstrating how the relationship between teacher quality and class size can affect the relative performance of the estimators.

Finally, this paper also relates to the literature viewing teacher payment as a contracting problem, where the administrator chooses the contract that induces the most effort given that she observes only a noisy measure of output (such as scores on standardized tests). Lazear and Rosen (1981) and Green and Stokey (1983) study optimal incentive schemes when worker effort is noisy and there may be aggregate shocks, finding that tournaments can provide proper incentives to workers. Hölmstrom and Milgrom (1991) focuses on providing incentives when an agent may split his effort on two types of tasks, only one of which admits a measure. Barlevy and Neal (2012) combines the earlier literature on tournaments and the multitask problem of Hölmstrom and Milgrom (1991) to specifically study the context of making comparisons between teachers. In contrast, in order to most closely match existing incentive schemes this paper assumes the administrator is following a cutoff rule, without proving that such a contract is optimal.[9] Therefore, this paper complements this literature by taking the distribution of teacher quality and choice of incentive scheme as given, and instead asks how commonly used estimators meet the administrator's objectives.

[9]Teacher risk-aversion will likely shape optimal incentive schemes based on high-stakes tests (Nadler and Wiswall (2011)), which the assumption of identifying teachers in a critical region likely rules out.

2 Model

A large body of empirical work evaluates VA models and compares the statistical properties of FE and EB estimators. However, determining which would be the *preferred* estimator for making decisions about rewarding or punishing teaches requires an economic model that posits an objective function for a decision maker, in this case, a school district administrator. To this end, I develop a simple model of the administrator's problem in this section.

The model formalizes the objective of a school-district administrator for one year; characterizes her optimal cutoff policy; and shows the relationship among i) how class size varies with teacher quality, ii) her choice of estimator, and iii) her expected utility. To most closely match observed bonus and/or retention policies, she takes as given an exogenous *desired cutoff* κ (for example, she is told to give bonuses to the top 5% quality teachers or to fire the lowest 1% quality teachers in the district) and chooses a *cutoff policy*, which varies by estimator type, to maximize her expected objective over all teachers in the district. My assumption that the administrator chooses estimator-specific cutoff policies differs from that in Schochet and Chiang (2012), who assume a fixed cutoff common to both estimators. The administrator receives utility from correctly rewarding a teacher with true quality equal to or higher than desired cutoff κ (not making a Type I error) and not rewarding a teacher with a true quality below κ (not making a Type II error). The administrator's utility from using the estimator $\hat{\theta}$ and cutoff policy c on a teacher of true quality θ is:

$$u(\theta, \hat{\theta}; c, \kappa) = \alpha \underbrace{\mathbf{1}\{\hat{\theta} \geq c \cap \theta \geq \kappa\}}_{\text{Not making Type I error}} + (1-\alpha) \underbrace{\mathbf{1}\{\hat{\theta} < c \cap \theta < \kappa\}}_{\text{Not making Type II error}},$$

where $\alpha, (1-\alpha)$ are her weights on not making Type I and II errors, respectively.

Teacher quality is distributed according to $\theta_i \sim F = \mathrm{N}\left(0, \sigma_\theta^2\right)$, where F is known.[10] The number of students assigned to teacher i, n_i, may depend on the teacher's quality. For example, a school principal may want to assign more students to a better teacher, or may assign the highest quality teachers to teach small groups of gifted students. For simplicity, I assume that class size is only a function of θ, where I sometimes explicitly denote this dependence by writing $n(\theta)$.[11] If class size were instead a noisy signal of teacher quality, the model solution would be more complicated, while not affecting which estimator the administrator would prefer. Moreover, I do not have data on the strength of principals' signals of teacher quality.

The test score gain for student j assigned to teacher i is $y_{ji} = \theta_i + \epsilon_{ji}$, where measurement error $\epsilon_{ji} \sim \mathrm{N}\left(0, \sigma_\epsilon^2\right)$ and $\epsilon_{ji} \perp\!\!\!\perp \theta_i$. I only adopt this simple technology for exposition in the current section; the VA estimates used in the quantitative portion of this paper control for many other characteristics. The fixed-effects (FE) estimator of θ_i is the sample mean, i.e., $\hat{\theta}_i^{FE} = \sum_j \frac{y_{ji}}{n_i} = \theta_i + \bar{\epsilon}_i$, and is distributed according to $\hat{\theta}_i^{FE} \sim \mathrm{N}\left(\theta_i, \frac{\sigma_\epsilon^2}{n_i}\right)$. The empirical Bayes (EB) estimator of teacher VA updates the prior (i.e. population) distribution of θ_i with data $\{y_{ji}\}_j$. Because both the prior distribution and measurement errors are assumed normal, the posterior distribution is also normal, giving $\hat{\theta}_i^{EB} = \lambda_i \hat{\theta}_i^{FE} + (1-\lambda_i) \underbrace{\mathrm{E}[\theta]}_{0} = \lambda_i \hat{\theta}_i^{FE} = \lambda_i(\theta_i + \bar{\epsilon}_i)$, where $\lambda_i = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\epsilon^2/n_i}$ is the ratio of the true variation in teacher quality (the signal) relative to the estimated variation using

[10] I follow the standard assumptions that teacher quality is normally distributed in the population, and that $\mathrm{E}[\theta]$ is normalized to 0 and estimated with infinite precision.

[11] If the number of students assigned to a teacher is a deterministic and strictly increasing function of teacher quality, teacher rankings could be recovered perfectly by comparing class sizes. Therefore, I assume that the administrator cannot perfectly recover quality rankings through class size rankings. In addition to complicating the scheme, which may reduce its attractiveness to policymakers, including $n(\theta)$ in the estimator would provide school principals with a direct incentive to manipulate class size, outside any effects of class size on total output.

the FE estimator (the signal plus noise).[12] Hereafter, I express the dependence of the weights on class size by writing $\lambda(n(\theta))$, $\lambda(n_i)$, or $\lambda(\theta)$, depending on which is more convenient. Note that how much the EB estimator is shifted towards the population mean depends on n_i; $\lambda(n_i) \to 1$ as the number of students observed for a teacher n_i increases, causing all the weight is shifted to the FE estimator. Also note that, though the EB estimate for a particular teacher's quality is biased ($\mathrm{E}\left[\hat{\theta}_i^{EB}\right] = \lambda(\theta)\theta_i \neq \theta_i$), the EB estimator for population teacher quality is unconditionally unbiased ($\mathrm{E}_\theta\left[\hat{\theta}_i^{EB}\right] = \mathrm{E}_\theta\left[\theta_i\right] = 0$).

Expected utility under the fixed effect estimator and cutoff policy c^{FE} integrates the administrator's objective over the distributions of teacher quality and measurement error:

$$\begin{aligned}
\mathrm{E}_{\theta,\bar{\epsilon}}\left[u(\theta,\hat{\theta}^{FE};c^{FE},\kappa)\right] &= \alpha\Pr\{\hat{\theta}^{FE} \geq c^{FE} \cap \theta \geq \kappa\} + (1-\alpha)\Pr\{\hat{\theta}^{FE} < c^{FE} \cap \theta < \kappa\} \\
&= \alpha\Pr\{\theta + \bar{\epsilon} \geq c^{FE} \cap \theta \geq \kappa\} + (1-\alpha)\Pr\{\theta + \bar{\epsilon} < c^{FE} \cap \theta < \kappa\} \\
&= \alpha\int_\kappa^\infty 1 - \Phi\left(\frac{c^{FE}-\theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) dF(\theta|\theta \geq \kappa) + (1-\alpha)\int_{-\infty}^\kappa \Phi\left(\frac{c^{FE}-\theta}{\sigma_{\bar{\epsilon}}(n(\theta)}\right) dF(\theta|\theta < \kappa), \quad (1)
\end{aligned}$$

where $\sigma_{\bar{\epsilon}}(n(\theta)) \equiv \frac{\sigma_\epsilon}{\sqrt{n(\theta)}}$ and $F(\theta|\theta \geq \kappa) = \frac{\phi(\theta/\sigma_\theta)}{\sigma_\theta(1-\Phi(\kappa/\sigma_\theta))}$ and $F(\theta|\theta < \kappa) = \frac{\phi(\theta/\sigma_\theta)}{\sigma_\theta\Phi(\kappa/\sigma_\theta)}$ are the distribution functions for θ, truncated below and above κ, respectively. Expected utility under the empirical Bayes estimator and cutoff policy c^{EB} is

$$\begin{aligned}
\mathrm{E}_{\theta,\bar{\epsilon}}\left[u(\theta,\hat{\theta}^{EB};c^{EB},\kappa)\right] &= \alpha\Pr\{\hat{\theta}^{EB} \geq c^{EB} \cap \theta \geq \kappa\} + (1-\alpha)\Pr\{\hat{\theta}^{EB} < c^{EB} \cap \theta < \kappa\} \\
&= \alpha\Pr\{\lambda(n(\theta))\hat{\theta}^{FE} \geq c^{EB} \cap \theta \geq \kappa\} + (1-\alpha)\Pr\{\lambda(n(\theta))\hat{\theta}^{FE} < c^{EB} \cap \theta < \kappa\} \\
&= \alpha\int_\kappa^\infty 1 - \Phi\left(\frac{\frac{c^{EB}}{\lambda(n(\theta))}-\theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) dF(\theta|\theta \geq \kappa) + (1-\alpha)\int_{-\infty}^\kappa \Phi\left(\frac{\frac{c^{EB}}{\lambda(n(\theta))}-\theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) dF(\theta|\theta < \kappa). \quad (2)
\end{aligned}$$

For either estimator, an increase in the cutoff policy c decreases the probability of correctly identifying a teacher with true quality above κ and increases the probability of correctly identifying a teacher with true quality below κ. The optimal cutoff solution equates the marginal increase in the probability of committing a Type I error (marginal cost) with the marginal decrease in the probability of committing a Type II error (marginal benefit). That is, c^{*EB} solves

$$\begin{aligned}
&\alpha\int_\kappa^\infty \frac{1}{\lambda(n(\theta))\sigma_{\bar{\epsilon}}(n(\theta))}\phi\left(\frac{c^{*EB}/\lambda(n(\theta))-\theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) dF(\theta|\theta \geq \kappa) \\
&= (1-\alpha)\int_{-\infty}^\kappa \frac{1}{\lambda(n(\theta))\sigma_{\bar{\epsilon}}(n(\theta))}\phi\left(\frac{c^{*EB}/\lambda(n(\theta))-\theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) dF(\theta|\theta < \kappa). \quad (3)
\end{aligned}$$

The optimal cutoff for the FE estimator c^{*FE} solves (3), where $\lambda(\theta) = 1, \forall\theta$. Denote the value to the administrator of using the optimal cutoff policies c^{*FE} and c^{*EB} as $v^{FE}(\kappa) = \mathrm{E}_{\theta,\bar{\epsilon}}\left[u(\theta,\hat{\theta}^{FE};c^{*FE},\kappa)\right]$ and $v^{EB}(\kappa) = \mathrm{E}_{\theta,\bar{\epsilon}}\left[u(\theta,\hat{\theta}^{EB};c^{*EB},\kappa)\right]$, respectively. The administrator's value for both estimators is increasing in the signal to noise ratio $\sigma_\theta/\sigma_\epsilon$: as the variance of the measurement error tends to 0, $\sigma_{\bar{\epsilon}} \to 0$, and all teachers will be correctly categorized, giving $v^{FE}(\kappa) = v^{EB}(\kappa) = 1, \forall\kappa$ (see Appendix B for the proof).

[12] McCaffrey et al. (2003) discusses the differences between FE and EB estimators. The EB estimator is pervasive in part because it minimizes mean squared error, i.e. it would be preferred by an administrator with a quadratic loss function. However, it is not clear that this is exactly how risk aversion should be modeled in this context. First, the administrator's loss function is not continuous in the model. Second, the administrator may be tasked with classifying an extremely large number of teachers, meaning that sampling variation in her classification errors may plausibly average out due to the weak law of large numbers. Finally, if there were only a small number of teachers being considered the population mean $\mathrm{E}[\theta]$ itself would not be precisely estimated, reducing the difference in variances between the FE and EB estimators.

2.1 Theoretical Results

I now characterize the administrator's value of using each estimator as a function of the relationship between teacher quality and class size. Proposition 1 shows that if there is no relationship between teacher quality and class size, the administrator's expected value is the same under both estimators. Next, I consider the case where class size depends on teacher quality. Proposition 2 shows that, in general, the administrator's expected value of the two estimators depends on the relationship between class size and teacher quality.

Proposition 1. *The administrator receives the same value from both estimators for any desired cutoff when class size is constant.*

Proof. If all classes are the same size then $\lambda(n(\theta)) = \lambda \in (0,1), \forall\theta$. Let c^{*FE} satisfy the administrator's first-order condition (3) when $\lambda = 1$. Because λ is constant, then $c^{*EB} = c^{*FE}\lambda$ also solves (3), and returns the same value ($v^{FE}(\kappa) = v^{EB}(\kappa)$); i.e. the maximized value of the administrator's objective is the same for both the FE and EB estimators. □

Figure 1 illustrates Proposition 1 by plotting the expected utility of the objective under the FE estimator (solid red line) and the EB estimator (dotted blue line) as a function of the cutoff policy for each estimator (x-axis), assuming the same class size for all teachers. Each curve traces out the administrator's expected utility as a function of cutoff policies, given an exogenous desired cutoff quality κ. The left panel corresponds to a desired cutoff of the first percentile teacher, i.e. $\kappa = F^{-1}(0.01)$, the middle panel corresponds to a desired cutoff of median teacher quality, i.e. $\kappa = F^{-1}(0.50)$ and the right panel corresponds to a desired cutoff of the 95th percentile teacher, i.e. $\kappa = F^{-1}(0.95)$. Extremely low or very high cutoff policies cause both estimators to misclassify either all low- or high-performing teachers, respectively, which is why the administrator's expected utility is 1/2 at either extreme. The utility-maximizing cutoff policy for each estimator is indicated by a vertical line $c^{*\text{estimator}}(\kappa)$, where the administrator's value from using that estimator, $v^{*\text{estimator}}(\kappa)$ is the maximum of each curve. Because the curves for both estimators obtain the same maximum height in each panel, we can see that these are equal when class size does not vary by teacher quality (i.e. $\lambda(n(\theta))$ is constant). The utility-maximizing cutoff policy adjusts to take into account the larger variance of administrator utility under the FE estimator. This is because if $c^{*FE} = \frac{c^{*FE}}{1}$ solves (3), $|c^{*EB}|$ must be smaller than $|c^{*FE}|$ to satisfy equation (3) if $\lambda < 1$ for the EB estimator. In the case where class size is constant, the optimal cutoff policies for both estimators are at the same quantiles of the estimator distributions, that is, the same share of teachers are rewarded under both estimators. For example, when the desired cutoff is the 1st percentile quality teacher, the cutoff policy that maximizes expected administrator utility when FE are used for inference is further to the left than when EB are used (Figure 1a). Alternatively, when the administrator desires to separate the top 5% (95th percentile) from the rest of teachers, the optimal cutoff policy under the FE estimator is higher than that under the EB estimator, again due to the larger variance of the FE estimator for teacher quality (Figure 1c). This higher variance does not affect the maximal value of the administrator's objective (maximum height of each curve), however, because the administrator is risk neutral.

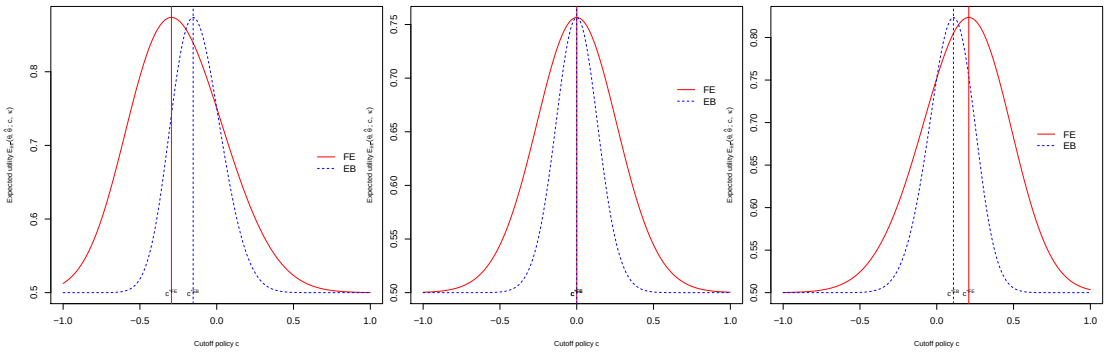
Proposition 2 considers the case where class size may depend on teacher quality.

Proposition 2. *In general, the administrator's preferred estimator depends on the relationship between teacher quality and class size.*

Proof. Because λ is increasing in n, to simplify the exposition I parametrize the EB weights λ directly as a function of θ, and then see how changes in this function would affect the administrator's utility from using the EB estimator. In particular, I assume there is one slope for the relationship between teacher quality and weight below the population mean, and another slope

Figure 1: Administrator's objective, assuming constant class size

(a) Desired cut 1% (b) Desired cut 50% (c) Desired cut 95%



for the relationship above the population mean. Furthermore, I set $\sigma_{\bar{\epsilon}} = 1$ for all teachers for the proof of the current proposition, which does not affect the qualitative result.[13] Let the reduced form for the EB weight be

$$\lambda(\theta) = \begin{cases} \delta_- + \beta_- \theta & \text{if } \theta < 0 \\ \delta_+ + \beta_+ \theta & \text{if } \theta \geq 0. \end{cases}$$

Suppose $\kappa < 0$, which implies $c^{*EB} < 0$. The result holds if $\kappa > 0$, using analogous reasoning. The administrator's value function is

$$\int_{-\infty}^{\kappa} \Phi\left(\frac{c^{*EB}}{\delta_- + \beta_- \theta} - \theta\right) dF(\theta|\theta < \kappa) + \int_{\kappa}^{0} \Phi\left(\theta - \frac{c^{*EB}}{\delta_- + \beta_- \theta}\right) dF(\theta|\theta \geq \kappa) + \int_{0}^{\infty} \Phi\left(\theta - \frac{c^{*EB}}{\delta_+ + \beta_+ \theta}\right) dF(\theta|\theta \geq \kappa). \tag{4}$$

Differentiate with respect to β_-:

$$\frac{\partial v}{\partial \beta_-} = \left[\int_{-\infty}^{\kappa} \frac{-c^{*EB}\theta}{(\delta_- + \beta_- \theta)^2} \phi\left(\frac{c^{*EB}}{\delta_- + \beta_- \theta} - \theta\right) dF(\theta|\theta < \kappa)\right] + \left[\int_{\kappa}^{0} \frac{c^{*EB}\theta}{(\delta_- + \beta_- \theta)^2} \phi\left(\frac{c^{*EB}}{\delta_- + \beta_- \theta} - \theta\right) dF(\theta|\theta \geq \kappa)\right],$$

where $\frac{\partial c^{*EB}}{\partial \beta_-} = 0$ due to the Envelope Theorem. The first term is positive because $-c^{*EB}\theta < 0$ for $\theta < \kappa$. Analogously, the second term is negative. Each term is the conditional mean of $\frac{c^{*EB}\theta}{(\delta_- + \beta_- \theta)^2}$, except weighted by the density $\phi\left(\frac{c^{*EB}}{\delta_- + \beta_- \theta} - \theta\right)$. Typically, the first term dominates, because it represents the conditional mean for the distribution of $\frac{c^{*EB}\theta}{(\delta_- + \beta_- \theta)^2}$ for the extreme part of the distribution of θ, while the second term is the conditional mean for the distribution of θ that is closer to the population mean of 0. If the first term dominates then the administrator's value is decreasing in β_-, i.e. the stronger the increase in class size from teacher quality. Analogously, by differentiating equation (4) with respect to β_+, we can see that the administrator's objective is increasing in β_+, meaning that increasing the weight associated with teacher fixed effects for teachers above the population mean improves the administrator's objective. Note that reducing the slope of class size in teacher quality for below-average teachers and increasing the slope of class size in teacher quality for above-average teachers improves the administrator's utility from using the EB estimator. In particular, if $\beta_- > 0$ and $\beta_+ < 0$, the FE estimator will provide the administrator with higher expected utility. □

[13] I allow $\sigma_{\bar{\epsilon}}$ to vary between teachers in the quantitative exercise.

Figure 2: Administrator's objective, assuming class size increasing in teacher quality

(a) Desired cut 1% (b) Desired cut 50% (c) Desired cut 95%

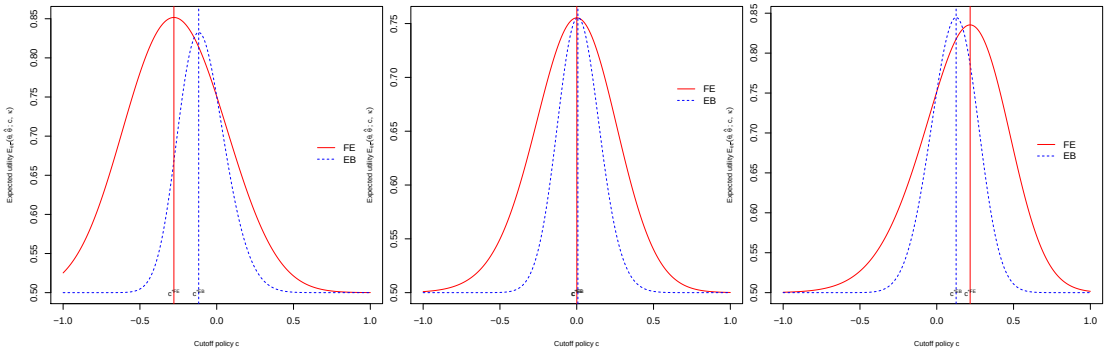
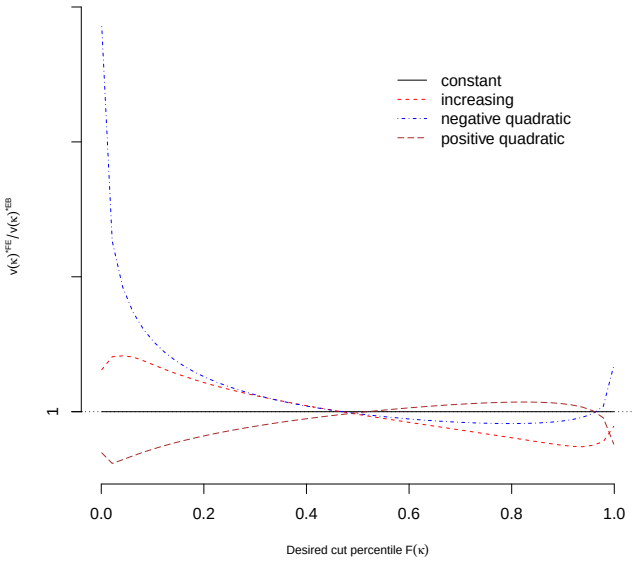


Figure 2 illustrates Proposition 2. Analogously to Figure 1, it plots the administrator's objective under both estimators ((1) and (2)) against cutoff policies (x-axis), but now under the assumption that class size is an increasing function of teacher quality. When class size is increasing in teacher quality, lower quality teachers are weighted closer to the population mean than higher quality teachers. If the administrator desires to separate the lowest quality teachers from the rest (Figure 2a), the re-weighting inherent in the EB estimator can actually reverse teacher rankings and lead to a lower expected objective for the administrator than when the FE estimator is used. The opposite is true for when the administrator wishes to separate the top teachers from the rest (Figure 2c) – the peak of the EB curve is now higher than that under the FE estimator. Intuitively, the EB estimator is now dilating the estimated teacher quality further than the FE estimator, reducing the probability the administrator makes a ranking error. When the administrator only desires to separate the upper and lower half quality teachers (Figure 2b), FE and EB both obtain the same maximum height, i.e. they return the same expected objective. An increase in either δ. corresponds to an increase in $\sigma_\theta/\sigma_\epsilon$, i.e. an increase in the signal to noise ratio. Intuitively, an increase in the signal provided by student test scores increases λ, reducing the dependence of the weight on teacher quality.

Figure 3 summarizes the theoretical results by comparing the performance of the estimators by plotting the ratio in value functions for the administrator ($v^{FE}(\kappa)/v^{EB}(\kappa)$) as a function of the desired cut percentile $F(\kappa)$ (x-axis), for scenarios where class size is constant, increasing in teacher quality, negative quadratic in teacher quality, and positive quadratic in teacher quality.[14] For each κ, estimator, and class size scenario, I solve for the administrator's optimal cutoff policy and plug it into her objective, returning $v^{\text{estimator}}(\kappa)$. The vertical axis then plots v^{FE}/v^{EB} corresponding to the desired cutoff associated with the desired cut percentile $F(\kappa)$. As shown before, when class size is constant (solid black line), the EB cutoff is just a scaled version of the FE cutoff and the administrator's value is the same under FE and EB estimators. When class size is increasing in teacher quality (short-dashed red line), the FE estimator performs better than the EB estimator when the administrator wishes to separate teachers of low quality from the rest (Figure 2a), while the EB estimator performs better when the administrator wishes to isolate high-quality teachers (Figure 2c). When class size has a negative quadratic relationship with teacher quality (dot-dashed blue line, similar to the case in Proposition 2 where $\beta_- > 0$ and $\beta_+ < 0$), it is

[14] The population average class size is the same for all four scenarios.

Figure 3: Difference between administrator's objective under FE and EB, by class size scenario and desired cut point



increasing when teacher quality is low and decreasing when teacher quality is high; in the example considered in Figure 3, the FE estimator outperforms the EB estimator at both the lowest and highest desired cutoffs. Finally, when class size is a positive quadratic function of teacher quality (long-dashed brown line), the opposite is true. Figure 3 also demonstrates that the difference between the performance of FE and EB estimators decreases the closer the desired cut point is to the population mean of 0. Intuitively, there is less of a difference between both the estimates resulting from the FE and EB estimators and the estimator-specific optimal cutoff policies when the administrator seeks to identify teachers as being on either side of the population mean (see Appendix C for the proof).

3 Quantitative Results

In this section, I quantify the relative value the administrator would receive from using the FE vs. EB estimators using data from the Los Angeles Unified School District, the second-largest school district in the US.[15] I assume the administrator wishes to categorize all teachers in the district with respect to an array of desired cutoffs in the district-wide distribution of teacher quality. Although such an incentive scheme is not currently in place in Los Angeles, this exercise serves as a useful benchmark for how the FE and EB estimators might perform when used in a similar incentive pay scheme.

The model shows that the difference in the administrator's value depends on the variances of teacher quality σ^2_θ and the test score measurement error σ^2_ϵ and the relationship between teacher quality and class size, $n(\theta)$, implying that it is necessary to obtain values for these objects to compare the performance of the estimators. Schochet and Chiang (2012) compile estimates of the

[15]Imberman and Lovenheim (2013) use these data in their study of the market's valuation of VA results.

variances from a large number of studies in their study of error rates in VA models, providing a good source for typical values for σ_θ^2 and σ_ϵ^2 (see Appendix D). The parameter values indicate that the variance of the measurement error is about 20 times the size of the variance of teacher quality, implying that student achievement for the average teacher in LA is about equal parts signal and noise. Therefore, it is difficult to correctly classify teachers. I obtain estimates of $n(\theta)$ from VA estimates provided by the LA Times, described below.

In 2011, the LA Times published the results of a RAND Corporation study estimating VA for 30,000 teachers serving almost 700,000 students (Buddin (2011)).[16] The dataset contains estimated VA for 3rd to 5th grade teachers in both reading and math and class sizes which condition on several variables, including past performance of students, student characteristics such as race, gender, English proficiency and parents education, and classroom composition (past performance of classmates and their student characteristics as well).[17] In addition to describing the relationship between teacher quality and class size, which is critical to compare the performance of the estimators, the distribution of VA estimates from the RAND study are comparable to those in Schochet and Chiang (2012).[18] The sample average class size is 22 students, with a standard deviation of 5 students.

Ideally, $n(\theta)$ would be known and fed into the administrator's problem. In practice, only estimates of $n(\theta)$, denoted by $\hat{n}(\hat{\theta})$, are available from the RAND study. Because of the point of this paper is to illustrate the difference in value when there is a very large number of teachers, the noisiness in this relationship is not of concern. Unfortunately, it is not clear which estimator was used in the data provided by the RAND study.[19] If FE were used in the RAND study, the shape (i.e. first and second derivatives) of the relationships between $n(\hat{\theta}^{FE})$ and $n(\theta)$ would be the same for all θ because the FE estimator is unbiased. However, it is potentially concerning if the estimates were computed using an EB estimator, given that the point of this paper is that differential shrinking may affect rankings of estimated VA, which may cause them to inaccurately characterize the relationship between true teacher quality and class size. I am able to show that the variation in class sizes in the data is not large enough to overturn the qualitative relationship between teacher quality and class size, even if biased EB estimates of teacher quality were used to recover it. Crucially, the calibrated values of error variances from Schochet and Chiang (2012) imply that λ_i is not variable enough to overturn the relationship I estimate between teacher quality and class size (see Appendix E for details).

The data show that, on average, teachers at either end of the distribution of reading VA have the smallest class sizes, while those in the middle of the distribution have the largest class sizes. [20] This pattern can be seen in Figure 4, which plots non-parametric regressions (solid blue lines) of class size on estimated teacher VA for reading (4a) and math (4b). Table 1 shows the results

[16] `http://projects.latimes.com/value-added/`

[17] The results do not qualitatively change when VA estimates from other specifications, which control for strict subsets of the above characteristics, are used instead.

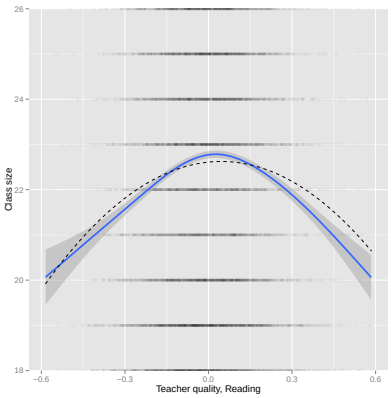
[18] The distribution of VA in the RAND data have means of 6.4E-11 and 1.3E-10 and variances of 0.038 and 0.083 for reading and math VA, respectively. Because the quantitative results combine data from the RAND study and parameter values calibrated from other datasets, the fact that these parameters are similar across the two types of sources lends validity to the quantitative results.

[19] Briggs and Domingue (2011) note that the methodology used to shrink VA estimates was not made clear in the technical report.

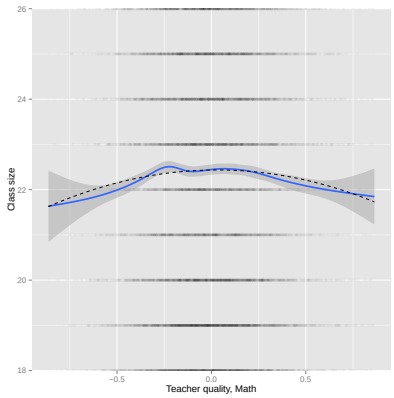
[20] I use annual data for class sizes, which is consistent with the view that a teacher's quality may change over the course of his/her career. Some authors have suggested using multiple years of student achievement growth to measure teacher VA (e.g. Kane and Staiger (2008), McCaffrey et al. (2009), Koedel and Betts (2011)). Although this would increase precision if teacher quality were unchanging, this would rule out any role of merit pay schemes to affect the productivity of current teachers, and further implies that changes in teacher quality can only be implemented by affecting sorting into and out of the profession of teaching. Even if it were appropriate to use data from multiple years, these simulation results apply to education systems that have only recently begun to collect and make available administrative data on student achievement. President Obama's Race to the Top education reform package incentivizes states to collect and make available results from standardized test scores.

Figure 4: The relationship between class size and teacher quality

(a) Reading



(b) Math



of regressions of teacher class size on teacher quality and teacher quality squared. The first two columns are for reading and the second two are for math. The dotted black lines on Figure 4 shows the regression line fit for models in columns (1) and (3). Columns (2) and (4) are the same as regressions in (1) and (3), respectively, but exclude teachers whose estimated quality is more than two standard deviations from the population mean. The results from this table indicate that class sizes are indeed increasing in VA in the lowest part of the distribution and decreasing in VA in the highest part of the distribution. The relationship is not as clear for math VA, but the regression shows that class size first increases and then decreases for reading VA, with a negative quadratic term for math VA. The results from this table indicate that class sizes are indeed increasing in VA in the lowest part of the distribution and decreasing in VA in the highest part of the distribution in most of the regressions. Strikingly, the observed relationship between teacher quality and class size is the worst case scenario for the EB estimator, as outlined by Proposition 2.

Figure 5 plots the ratio of the administrator's maximized expected objective under the FE and EB estimators for Reading (solid black line) and Math (dotted red line), using the calibrated parameter values and estimated relationship between teacher quality and class size. The left panel (5a) plots the gain in expected value ($v^{FE}(\kappa)/v^{EB}(\kappa)$) to switching from the EB to the FE estimator, for cutoff percentiles ranging from the lowest to the highest teacher qualities. The right panel (5b) plots how many more expected mistakes the EB estimator would make than the FE estimator, assuming the Los Angeles school district employed 30,000 teachers.[21] We can see that the quadratic nature of the association between teacher quality and class size affects the relative performance of the FE and EB estimators in the way demonstrated by Proposition 2. The stronger negative quadratic relationship between teacher quality and class size in the Reading test causes the larger divergence between the expected value of using FE rather than EB estimators. The administrator's value function is higher almost everywhere when she uses the FE estimator of teacher VA, and the relative performance of the EB estimator is the worst at the extremes of the distribution of teacher quality, where using FE gives the administrator 1% higher expected objective, corresponding to the EB estimator making over 200 more mistakes than the FE estimator when Reading scores are used. The administrator's expected values from using the FE and EB estimators are comparable as the desired cutoff approaches the center of the distribution of teacher quality, as was the case in Figure 2, where class size was increasing in teacher quality. In

[21]The Los Angeles school district is the second-largest in the US. Though the VA data I am using cover 30,000 teachers, more than 45,000 worked in the district in 2007 (`http://en.wikipedia.org/wiki/Los_Angeles_Unified_School_District`).

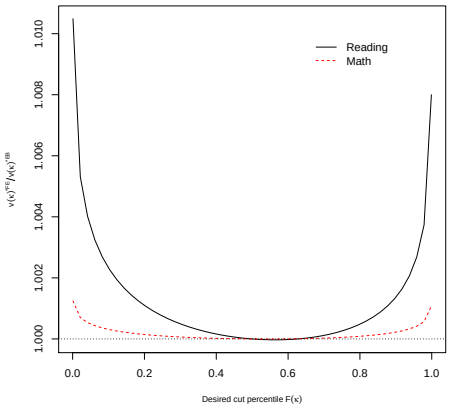
Table 1: Regressions of class size on teacher quality

	Dependent variable: Class size			
	(1)	(2)	(3)	(4)
Reading quality	0.618^{***}	0.650^{***}		
	(0.139)	(0.167)		
Sq. Reading quality	-6.801^{***}	-11.180^{***}		
	(0.368)	(0.834)		
Math quality			0.060	-0.008
			(0.092)	(0.109)
Sq. Math quality			-1.014^{***}	-1.527^{***}
			(0.212)	(0.370)
Constant	22.609^{***}	22.736^{***}	22.434^{***}	22.467^{***}
	(0.030)	(0.035)	(0.032)	(0.035)
Observations	36,125	34,407	36,125	34,372
R^2	0.009	0.006	0.001	0.0005
F Statistic	170.442^{***}	99.271^{***}	11.442^{***}	8.535^{***}
	(df = 2; 36122)	(df = 2; 34404)	(df = 2; 36122)	(df = 2; 34369)

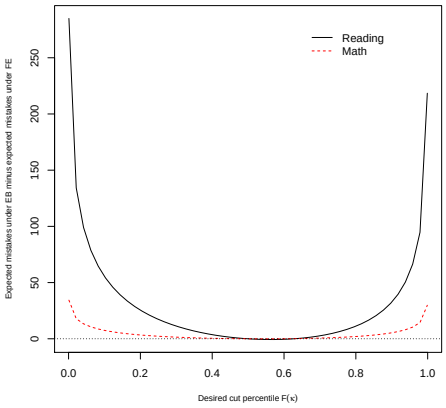
***p<0.01

Figure 5: Administrator's value from using FE and EB, using $n(\theta)$ from RAND study

(a) Probability of correct classification (v^{FE}/v^{EB})



(b) Expected number of mistakes (EB - FE)



sum, the performance of the FE and EB estimators most greatly diverges precisely where policies that sanction very low performing teachers or reward very high performance teachers would bite the most, and the FE estimator returns higher expected utility (i.e. makes fewer mistakes, in expectation) under almost every desired cutoff.

4 Conclusion

Society must use an economic, not statistical, rationale when forming education policy. At the same time, measurement issues are important when considering how estimators based on data from finite samples would perform when part of actual educational policies. Empirical Bayes estimators of teacher value-added are used by many education researchers and practitioners to make inferences about teacher quality, which may serve as inputs to high-stakes decisions like bonus assignments or teacher personnel decisions. In this paper, I show that the value of EB vs FE estimators depends critically on the relationship between class size and teacher quality. My theoretical findings suggest that an inverted U-shaped relationship between teacher quality and class size leads the FE estimator to outperform the EB estimator, because the latter re-weights teachers with quality at either extreme closer to the average quality in the population, making it difficult to identify them relative to other teachers. Finally, I show that class size and teacher quality are inverted-U shaped in the Los Angeles Unified School District, the second largest school district in the US. For almost every desired cutoff an administrator might choose in a prospective incentive scheme, this would cause the FE estimator to outperform the EB estimator. At the extremes, the EB estimator would make more than 200 more classification mistakes than the FE estimator when classifying teachers based on their students' Reading test scores.

Although this paper characterizes when the FE estimator should be preferred by a risk-neutral administrator and quantifies their performance in an extremely large school district that has recently received much policy interest (such as that created by the LA Times release of RAND VA estimates used here), a more comprehensive study of when and where FE estimators should be preferred to EB estimators would require data from the relevant geography and information about the administrator's preferences. Additionally, this paper is silent about the use of VA as a measure of teacher quality, which means that even though the FE estimator may outperform the EB

estimator within the class of VA models, neither may perform very well if the true technology were not consistent with a VA model. That being said, this paper demonstrates the importance of specifying an economic model of policymaker utility, which provides guidance about which estimator would be preferred.

A Teacher incentive schemes

Table 2: Incentive pay schemes based in part on student achievement

Name of scheme	Location	Active dates	Bonus schedule	Uses VA?	Uses EB?
Dallas Independent School District (DISD) Principal and Teacher Incentive Pay program	Dallas, Texas	2007-08 school year (Previous program started in 1992)	Discrete	Yes	Yes
TVAAS	Tennessee	Since 1996	Discrete	Yes	Yes
Tennessee Educator Acceleration Model (TEAM)	Tennessee	Since 2010	Discrete	Yes	Yes
Memphis' Teacher Effectiveness Measure (TEM)	Memphis, Tennessee	Since 2010	Discrete	Yes	Yes
Pennsylvania	Pennsylvania	Since 2013-2014	Discrete	Yes	Yes
Pittsburgh	Pittsburgh	Since 2013-2014	Discrete	Yes	Yes
North Carolina Teacher Evaluation Process	North Carolina	since 2012-2013	Discrete	Yes	Yes
Mission Possible	Guilford County, North Carolina	2006-current	Discrete	Yes	Yes
Milken Family Foundation's Teacher Advancement Program (TAP)	Nationwide (125 schools in 9 states and 50 districts as of 2007)	Since 1999	Discrete	Yes	Varies
Denver Public School's Professional Compensation System for Teachers (ProComp)	Denver, Colorado	Since 2005	Discrete (many bonus levels)	No	No
Special Teachers Are Rewarded (STAR) (followed by MAP)	Florida	2006-2007 (MAP since 2007)	Discrete (MAP has both continuous and discrete rewards)	No (though they do use a discretized version of VA through a value table)	No
North Carolina ABCs	North Carolina	1996-2012	Discrete	No	No
Q-Comp	Minnesota	Since 2005	Varies, but mostly discrete	Varies between participants, but unknown in general.	?
Louisiana	Louisiana	Since 2010	Discrete	?	?
Texas' Governor's Educator Excellence Award Programs (GEEAP)	Texas	2008 school year	?	?	?

Source: Author's compilation. References available upon request.

B Infinite precision

We want to prove that as the variance of the measurement error tends to 0 (which implies $\sigma_{\bar{\epsilon}} \to 0$) all teachers will be correctly categorized, giving $v^{FE} = v^{EB} = 1$. First consider the FE estimator. The administrator's utility for a teacher with true quality θ under estimator $\hat{\theta}$ and cutoff policy c is

$$u(\theta, \hat{\theta}; c, \kappa) = \alpha \mathbf{1}\{\hat{\theta} \geq c \cap \theta \geq \kappa\} + (1-\alpha)\mathbf{1}\{\hat{\theta} < c \cap \theta < \kappa\} \quad \xrightarrow{p} \alpha \mathbf{1}\{\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)\mathbf{1}\{\theta < c \cap \theta < \kappa\}, \tag{5}$$

which is maximized at $c = \kappa$. Note that the administrator's utility from using EB estimator for the same teacher is

$$\begin{aligned} plim_{\sigma_{\bar{\epsilon}} \to 0} u(\theta, \hat{\theta}; c, \kappa) &= \alpha \mathbf{1}\{\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)\mathbf{1}\{\theta < c \cap \theta < \kappa\} \\ &= \alpha \mathbf{1}\{\lambda(\theta)\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)\mathbf{1}\{\lambda(\theta)\theta < c \cap \theta < \kappa\}, \end{aligned} \tag{6}$$

which is maximized at $c = \kappa/\lambda(F^{-1}(\kappa))$.[22] The probabilities of the events in both (5) and (6) are all 1, giving an expected utility of 1 for all teacher qualities, which then integrates to a value of 1 for each estimator.

C Proof that FE and EB give equivalent expected utility when the administrator's problem is symmetric

Definition 1. *The administrator's problem is symmetric if $n(\theta)$ is symmetric around the population mean of teacher quality and the administrator's desired cutoff $\kappa = 0$.*

Proposition 3. *The administrator receives the same value from both estimators when the problem is symmetric.*

Proof. Because $n(\theta)$ is symmetric about $\theta = 0$ and $\theta_i \sim F = \mathrm{N}\left(0, \sigma_\theta^2\right)$, the distribution of θ is symmetric around its population mean of 0. If $\kappa = \mathrm{E}\left[\theta\right] = 0$, and $\alpha = 1/2$, the optimal c^{*EB} solves

$$\int_0^\infty \frac{1}{\lambda(n(\theta))\sigma_{\bar{\epsilon}}(n(\theta))} \phi\left(\frac{c^{*EB}/\lambda(n(\theta)) - \theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) \phi(\theta/\sigma_\theta) d\theta = \int_{-\infty}^0 \frac{1}{\lambda(n(\theta))\sigma_{\bar{\epsilon}}(n(\theta))} \phi\left(\frac{c^{*EB}/\lambda(n(\theta)) - \theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) \phi(\theta/\sigma_\theta) d\theta.$$

At $c^{*EB} = 0$, the expression becomes

$$\int_0^\infty \frac{1}{\lambda(n(\theta))\sigma_{\bar{\epsilon}}(n(\theta))} \phi\left(\frac{-\theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) \phi(\theta/\sigma_\theta) d\theta = \int_{-\infty}^0 \frac{1}{\lambda(n(\theta))\sigma_{\bar{\epsilon}}(n(\theta))} \phi\left(\frac{-\theta}{\sigma_{\bar{\epsilon}}(n(\theta))}\right) \phi(\theta/\sigma_\theta) d\theta,$$

which holds because of the symmetry of $\phi()$, $n()$, and $\lambda()$ (through its dependence on n, which is symmetric). Therefore, $c^{*EB} = 0$ solves the administrator's problem when EB is used. Because $\lambda(n(\theta)) = 1$, $\forall\theta$ when the FE estimator is used, $c^{*FE} = 0$ must also solve the administrator's problem when FE is used, and the administrator's objective is equivalent under both estimators. □

D Calibrated parameters

I calibrate σ_θ^2 and σ_ϵ^2 from Schochet and Chiang (2012), Table B-2. To most closely match a policy where an administrator would like to rank teachers across an entire school district, I calibrate $\sigma_\theta^2 = 0.046$ by summing the average of school- and teacher-level variances in random effects. To

[22] Recall that $\theta_i \sim F = \mathrm{N}\left(0, \sigma_\theta^2\right)$.

most closely approximate an environment where both student and aggregate-level shocks may affect student test scores, I calibrate $\sigma^2_\epsilon = 0.953$ by summing the average of class- and student-level variances in random effects. Note that, due to the vastly greater student-level error variance, the approximate sizes of σ^2_θ and σ^2_ϵ are approximately the same if school-level variances are excluded from σ^2_θ or class-level variances are excluded from σ^2_ϵ, lending robustness to the quantitative findings.

E Proof that the relationship between teacher quality and class size is robust to the (potential) use of EB estimators in the RAND study

We observe $\hat{n}(\hat{\theta})$ and want to make an inference about $n(\theta)$. Because I am focusing on the qualitative relationship between teacher quality and class size, I consider the derivative for various true θ, and examine when the true and estimated relationships have the same sign.

Differentiate the observed class size at θ_0:

$$\frac{\partial \hat{n}}{\partial \theta} = \frac{\partial n(\theta_0)}{\partial \theta} \times \left[\frac{\partial \lambda(n(\theta_0))}{\partial n} \frac{\partial n(\theta_0)}{\partial \theta} (\theta_0 + \bar{\epsilon}) + \lambda(n(\theta_0)) \right],$$

and take expectations with respect to the measurement error because we are interested in characterizing the average relationship of class size given teacher quality:

$$\underbrace{\frac{\partial \hat{n}}{\partial \theta}}_{?} = \underbrace{\frac{\partial n(\theta_0)}{\partial \theta}}_{?} \times \left[\underbrace{\frac{\partial \lambda(n(\theta_0))}{\partial n}}_{>0} \underbrace{\frac{\partial n(\theta_0)}{\partial \theta}}_{?} \underbrace{\theta_0}_{<0 \cap \geq 0} + \underbrace{\lambda(n(\theta_0))}_{>0} \right], \tag{7}$$

which shows that $\hat{n}'$ and n' have the same sign as long as the term in square brackets is positive. The FE estimator sets $\lambda_i = 1, \forall i$ which means $\frac{\partial \lambda(n(\theta_0))}{\partial n} = 0, \forall \theta_0$. Therefore, $\hat{n}'$ and n' always have the same sign if the FE estimator had been used. Table 3 outlines the four scenarios that arise when using the EB estimator.

Table 3: Scenarios relating teacher quality and class size

	Condition	
	$n'(\theta_0) > 0$	$n'(\theta_0) < 0$
$\theta_0 > 0$	I	II
$\theta_0 < 0$	III	IV

First, note that the signs of $\hat{n}'$ and n' match in scenarios I and IV. Algebraic manipulation shows their signs also match in scenarios II and III so long as

$$\left| \frac{\partial n(\theta_0)}{\partial \theta} \right| < \frac{\lambda(n(\theta_0))}{\frac{\partial \lambda(n(\theta_0))}{\partial n} \times |\theta_0|}.$$

I use $n = 11$ (the smallest class size in the sample) as a conservative measure of $n(\theta_0)$, because the right hand side is increasing in n. Evaluating the expression for the 1st percentile teacher quality, the left hand side must be smaller than 33.75, which is associated with a change of more than 7 students when teacher quality moves by one standard deviation - well outside the range

observed in the data for the average class size for an estimated teacher quality. Therefore, though the fact that the estimates *may* be EB may attenuate the relationship between class size and teacher quality that feeds into the model, we can be confident that the results of the empirical portion of this paper are qualitatively correct.

References

Andrabi, T., et al. Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal: Applied Economics*, 3(3):29–54 (2011).

Baker, E. and Barton, P. Problems with the Use of Student Test Scores to Evaluate Teachers. *Economic Policy Institute*, EPI Briefing Paper #278. (2010).

Barlevy, G. and Neal, D. Pay for percentile. *American Economic Review* (2012).

Bond, T. N. and Lang, K. The evolution of the black-white test score gap in grades k–3: The fragility of results. *Review of Economics and Statistics*, 95(5):1468–1479 (2013).

Briggs, D. and Domingue, B. Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of los angeles unified school district teachers by the "los angeles times". *National Education Policy Center* (2011).

Buddin, R. Measuring teacher and school effectiveness at improving student achievement in los angeles elementary schools. *RAND Corporation Working paper* (2011).

Chetty, R., et al. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates (2013).

Ferrall, C. and Shearer, B. Incentives and transactions costs within the firm: estimating an agency model using payroll records. *The Review of Economic Studies*, 66(2):309–338 (1999).

Glazerman, S., et al. Evaluating teachers: The important role of value-added. Technical report, Mathematica Policy Research (2010).

Glazerman, S., et al. Impacts of performance pay under the teacher incentive fund: Study design report. *Mathematica Policy Research, Inc.* (2011).

Goldhaber, D. D. and Brewer, D. J. Why don't schools and teachers seem to matter? assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, 32(3) (1997).

Goldstein, D. Randi weingarten: Stop the testing obsession. *Dana Goldstein's Blog at The Nation* (2012).

Green, J. and Stokey, N. A comparison of tournaments and contracts. *The Journal of Political Economy*, 91(3):349–364 (1983).

Guarino, C., et al. Can value-added measures of teacher performance be trusted? (2012a).

Guarino, C. M., et al. An evaluation of empirical bayes estimation of value-added teacher performance measures. *Education Policy Center at Michigan State University Working Paper*, 31 (2012b).

Hanushek, E. Conceptual and empirical issues in the estimation of educational production functions. *Journal of human Resources*, pages 351–388 (1979).

Hanushek, E. The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24(3):1141–1177 (1986). ISSN 0022-0515.

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

Hölmstrom, B. and Milgrom, P. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.* (1991).

Imberman, S. A. and Lovenheim, M. F. Does the market value value-added? evidence from housing prices after a public release of school and teacher value-added. Technical report, National Bureau of Economic Research (2013).

Jacob, B. and Lefgren, L. Principals as agents: Subjective performance measurement in education. *NBER Working Paper*, No. 11463 (2005).

Kane, T. J. and Staiger, D. O. Estimating teacher impacts on student achievement: An experimental evaluation. *NBER Working Paper* (2008).

Kane, T. J., et al. What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education Review*, 27(6):615–631 (2008).

Kinsler, J. Assessing rothstein's critique of teacher value-added models. *Quantitative Economics*, 3(2):333–362 (2012).

Koedel, C. and Betts, J. Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education* (2011).

Lazear, E. Educational production. *Quarterly Journal of Economics*, pages 777–803 (2001).

Lazear, E. and Rosen, S. Rank-order tournaments as optimum labor contracts. *The Journal of Political Economy* (1981).

McCaffrey, D. F., et al. *Evaluating Value-Added Models for Teacher Accountability. Monograph.* ERIC (2003).

McCaffrey, D. F., et al. The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, 4(4):572–606 (2009).

McGuinn, P. Stimulating reform race to the top, competitive grants and the obama education agenda. *Educational Policy*, 26(1):136–159 (2012).

Nadler, C. and Wiswall, M. Risk aversion and support for merit pay: Theory and evidence from minnesota's q comp program. *Education Finance and Policy*, 6(1):75–104 (2011).

Podgursky, M. and Springer, M. Teacher performance pay: A review. *National Center on Performance Incentives*, pages 2006–01 (2006).

Podgursky, M. and Springer, M. Teacher compensation systems in the united states k-12 public school system. *National Tax Journal*, 64(1):165–192 (2011).

Rivkin, S., et al. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458 (2005). ISSN 1468-0262.

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

Rothstein, J. Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, 4(4):537–571 (2009).

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

Schochet, P. and Chiang, H. What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics* (2012).

Stiglitz, J. E. Symposium on organizations and economics. *The journal of economic perspectives*, pages 15–24 (1991).

Strauss, V. Errors found in d.c. teacher evaluations. *The Washington Post* (2013).

Tate, R. A cautionary note on shrinkage estimates of school and teacher effects. *Florida J. Educ. Res*, 42:1–21 (2004).

Todd, P. and Wolpin, K. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485):F3–F33 (2003).

Turque, B. Rhee dismisses 241 d.c. teachers; union vows to contest firings. *The Washington Post* (2010).