Measuring Quality for Use in Incentive Schemes:

The Case of "Shrinkage" Estimators

Nirav Mehta
University of Western Ontario *

April 2, 2019

Abstract

Researchers commonly "shrink" raw quality measures based on statistical criteria. This paper studies when and how this transformation's statistical properties would confer economic benefits to a utility-maximizing decisionmaker across common asymmetric information environments. I develop the results for an application measuring teacher quality. The presence of a systematic relationship between teacher quality and class size could cause the data transformation to do either worse or better than the untransformed data. I use data from Los Angeles to confirm the presence of such a relationship and show that the simpler raw measure would outperform the one most commonly used in teacher incentive schemes.

Keywords: economics of education, empirical contracts, teacher incentive schemes, teacher quality

JEL codes: J01, I21, I28, D81

*nirav.mehta@uwo.ca I thank Tim Conley, Steven Glazerman, Dan Goldhaber, Rick Hanushek, Lance Lochner, Rachel Margolis, Henry May, Seth Richards-Shubik, Gil Shapira, Todd Stinebrickner, and Michela Tincani for useful discussions pertaining to this paper, and the SSHRC Insight Development Grant program and Jacobs Foundation for funding. I also thank Enrique Martin Luccioni for research assistance.

1 Introduction

In response to what is thought to be excessive noise present in directly observed measures of important economic inputs (e.g., teacher quality), many researchers and practitioners transform raw measures by "shrinking" them towards the population mean. Shrinking the raw measure by a factor decreasing in the number of observations used to compute it results in a "shrinkage estimator", which minimizes mean squared error, making it the best predictor. This well-known statistical property (Copas (1983); Morris (1983)) has motivated the use of shrinkage estimators to inform decisions in a large number of policy-relevant applications, including (but not limited to) the estimation of teacher quality (Rockoff (2004); Kane et al. (2008); Chetty et al. (2014a,b)), school quality (Raudenbush and Bryk (1986); Angrist et al. (2017)), and neighborhood effects (Oakes (2004); Chetty and Hendren (2016)). Shrinkage estimators have also been used to set insurance premia (Makov et al. (1996)) and, more generally, in the evaluation of social programs (Rossi et al. (2003)).

The starting point for this paper is to recognize that quality measures are rarely of intrinsic value, but rather, they are important because they are used to make decisions. Statistically desirable properties may not confer advantages when viewed from an economic perspective, i.e., when used by an optimizing decisionmaker in the relevant context. For example, consider how to assign a bonus to the best teacher in a school, where the optimal policy based on raw data would reward the teacher with the highest measured output. If the best teacher had a relatively small classroom (i.e., fewer observations) and therefore was shrunk closer to the mean, shrinking the raw measure could lower the odds of rewarding the best teacher. It is not clear that an estimator with a lower mean squared error should always be preferred to an unbiased one.

This paper examines whether a utility-maximizing decisionmaker in an asymmetric information environment could improve upon the shrinkage estimators pervasive in research and practice. I first examine a practical and concrete cutoff model, where the decisionmaker classifies agents with respect to a desired threshold to minimize a weighted sum of the expected Type I and Type II errors. To study how measurement affects output, I next study hidden type (adverse selection) and hidden action (moral hazard) models, where agent quality is, respectively, fixed and endogenous. Optimal policy in the hidden type model is a stopping rule, or reservation quality measure, which suggests an economic intuition for the cutoff model. Optimal policy in the hidden action model, which is based on Hölmstrom and Milgrom (1987), is linear in the quality measure.

In each environment, the decisionmaker chooses the best estimator from a set containing a (perhaps) naive measure, the raw, or unshrunken, quality measure, and the ubiquitous

shrunken measure by comparing the value—that is, maximized expected utility obtained under optimal policy—according to each estimator. Estimators in this set yield the clearest insights because they only differ by how much they weigh sample data, which increases in the number of observations, or sample size, per agent. For example, in the context of estimating teacher quality these weights depend on class size; if class size systematically differed between teachers then the amount of shrinkage could be related to underlying quality.[1]

The theoretical analysis shows that taking into account the decisionmaker's optimization behavior can undo or even reverse an estimator's statistical advantages. The main theoretical result—that the relationship between sample size and quality determines the preferred estimator—is common across environments. When sample sizes (and thus, shrinkage) are constant, optimal policy would undo any shrinkage, *eliminating the desirable statistical properties of shrinkage estimators* and leaving the decisionmaker indifferent between estimators. Nonconstant sample sizes that are related to quality result in "differential shrinkage", creating a difference in the value according to each estimator. For example, when the sample size is negative-quadratic in quality, the decisionmaker may prefer the raw quality measure and when it is positive-quadratic in quality she may prefer the shrunken one.

I develop the analysis using a highly policy-relevant application: the measurement of teacher quality. The fact that only a small amount of variation in student achievement is explained by teachers' observed characteristics that are available in typical administrative or survey data (Hanushek (1986); Rivkin et al. (2005)) and evidence that teacher quality is an important determinant of human capital (Hanushek (2011); Chetty et al. (2014a)) have spurred the introduction of teacher incentive schemes. For example, President Obama's Race to the Top initiative incentivizes states to adopt incentive pay schemes and the Teacher Advancement Program has introduced performance-based bonuses to over 20,000 teachers serving over 200,000 students across the U.S.[2] In addition to being a linchpin of education reform, teacher incentive schemes may be the most visible incarnation of performance-based incentives in the public sector. The concern that teacher quality measures can be quite noisy (Baker and Barton (2010)), which may subject teachers to undue risk if directly used in high-powered incentive schemes, and the fact that they minimize mean squared error has motivated the use of shrinkage estimators—most commonly "empirical Bayes"—in this application.[3]

[1] As discussed below, there is a large literature studying this relationship. This paper's contribution is to show how such a relationship could interact with measurement from the perspective of a utility maximizer.

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

[3] 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)).

The theoretical environments described above are relevant for studying teacher quality, where the decisionmaker could be a school district administrator, i.e., a public official interested in cost-effective ways of increasing educational production. As I document in Appendix A, the cutoff model matches the structure of the vast majority of existing teacher incentive schemes, which typically use empirical Bayes to measure teacher quality when assigning bonuses or even dismissing teachers; it could also be used to model pay-for-percentile-type schemes, which are tournament-based schemes that have recently become popular in education policy debates (Barlevy and Neal (2012)).[4] Optimal policies in the asymmetric information models take on the natural interpretation of reward, or wage, schedules that weakly increase in measured quality. The optimal stopping rule in the hidden type model is a step function, where teachers with above-threshold measures receive the same (positive) salary and those below receive a wage of zero (meaning they are dismissed). The optimal wage schedule in the hidden action model is a constant base salary, with a performance-based bonus that increases linearly in measured quality. Notably, in each environment, the administrator chooses an optimal reward policy function for each estimator. This is in contrast to the modal approach, which quantifies the effects of potentially suboptimal policies using estimated models (e.g., Stinebrickner (2001), Tincani (2012), Todd and Wolpin (2012), Behrman et al. (2016)) or those with calibrated parameters (e.g., Rothstein (2014)). While clearly attractive, characterizing optimal reward policies is only possible for environments that are parsimonious relative to those typically considered in this literature. Viewed in light of this tradeoff and their somewhat different goals, this paper is quite complementary to this literature.

The aforementioned concern that the amount of shrinkage could be systematically related to the underlying estimand is germane to measuring teacher quality. The idea that class size can reflect information about teacher quality has theoretical precedent (Lazear (2001); Barrett and Toma (2013)) and researchers have found that higher-quality teachers tend to have more favorable working conditions, in terms of student characteristics (Player (2010), Clotfelter et al. (2006)); it is not a big leap to extend this reasoning to class size.[5]

When measuring teacher quality, the administrator's preferred estimator is determined by the relationship between class size and teacher quality. In particular, she would prefer the

[4]The cutoff model also allows us to be agnostic about what underlies variation in measured output, which could be due to hidden types and/or actions.

[5] Barrett and Toma (2013) assume that higher-quality teachers have a smaller reduction in efficacy for a given increase in class size, which could induce a relationship between class size and teacher quality. Jepsen and Rivkin (2002, 2009) show that a funding increase resulted in smaller class sizes, though teachers hired to affect this reduction had less experience. Because these teachers were likely far less effective than experienced ones, this finding would be consistent with a positive relationship between class size and teacher quality at the low end of the distribution.

same estimator across the three models: she would be indifferent when class size was constant in teacher quality, prefer fixed effects when class size was negative-quadratic in teacher quality, and prefer empirical Bayes when class size was positive-quadratic in teacher quality. This finding is very intuitive. As discussed above, with constant class sizes her optimal policy in each model would adjust to return the same estimator-specific value. Roughly, in the cutoff and hidden type models, the administrator would prefer fixed effects when the teachers with qualities she would most like to identify are those with relatively small class sizes. Because a negative-quadratic relationship between class size and teacher quality would leave teachers at the low and high ends of the quality distribution with the smallest class sizes, this would lead to a lower value from empirical Bayes in the cutoff model for a wide range of desired cutoffs, as shrinking makes it harder to sanction/fire less-effective teachers or reward/retain more-effective teachers. Similarly, in the hidden type model, the administrator's optimal policy is to try to identify teachers with qualities below a certain level, which is harder to do when shrinking the data when class sizes for these teachers are small. Finally, a negative-quadratic relationship between class size and teacher quality decreases signal precision of the shrunken quality measure, reducing the resulting optimal incentive strength and optimized output in the hidden action model, again leading the administrator to prefer fixed effects in the presence of a negative-quadratic relationship between class size and teacher quality.

The empirical part of the application uses data from the Los Angeles Unified School District, the second-largest school district in the U.S. and one with a large degree of diversity and variation in both student achievement and class size (Buddin (2011)). I find that class sizes are smallest for the lowest- and highest-quality teachers, which is the scenario in which the raw quality measure would be preferred to the shrunken measure in each environment. There is reason to also expect the type of relationship between class size and teacher quality I document in Los Angeles in other locales. Suppose that, in the background of the administrator's optimization problem, school principals wanted to have students pass a low proficiency threshold and increase total output at their respective schools. The former could cause class size to increase in teacher quality at the low end of the quality distribution. However, due to the lack of flexible wages in the public education sector, school principals might also reduce class size at the high end of the distribution to retain high-quality teachers.[6] This paper remains agnostic about the source of this relationship; all that matters is whether it exists and, if it does exist, what it is.

Finally, I calibrate additional parameters to quantitatively compare the prospective per-

[6] This example provides only one potential explanation underlying the documented relationship between class size and teacher quality, which would clearly be constrained by institutional factors like teacher unions and, relevant for my empirical application, the California Education Code, Section 41376, which prescribes maximum class sizes and penalties for districts violating those limits.

formance of the estimators, finding nontrivial benefits to using raw quality measures. For example, in the cutoff model, an administrator would make 9% more classification errors when using empirical Bayes when seeking to identify teachers in the bottom 1% in Reading value-added and switching from empirical Bayes would increase output by 2% in the hidden action model. The performance of the estimators in the cutoff model differs most at the tails of the distribution of teacher quality, which is important for identifying either high- or low-quality teachers, the focus of existing teacher incentive schemes (e.g., the Washington D.C. Schools Chancellor fired 241 teachers in 2010 based on performance measures (Turque (2010)).[7] The sheer number of schemes and affected teachers and students and increasing policy support for teacher incentive schemes point to substantial gains from using the preferred estimator for the relevant context, especially in light of the relatively costless "intervention" of adopting a simpler quality measure. Because incentive schemes do not directly take into account class size, when there is a relationship between class size and teacher quality the choice of estimator can effectively help to take into account this relationship.

Section 2 presents statistical background for the models used in this paper. Section 3 develops and analyzes the cutoff model and Section 4 presents the hidden type and hidden action models. Section 5 presents the quantitative results and Section 6 concludes. The Appendix documents a number of teacher incentive schemes; contains proofs and further details about the quantitative results are in the Online Appendix.

2 Statistical Background

The application to measure teacher quality is based on the leading conceptual framework for teacher quality, the value-added model (Murnane (1975); Hanushek (1979)), which uses changes in students' test scores over the year to measure the contribution (i.e., quality) of individual teachers. There is a literature studying the statistical properties of the value-added framework, with the main concern that the omission of important inputs may bias estimates (e.g., McCaffrey et al. (2003), Rothstein (2009, 2010), Glazerman et al. (2010), Andrabi et al. (2011), Guarino et al. (2014), Jackson (2014)). However, several recent studies have found that value-added models are fairly good at accounting for unobserved inputs (Kinsler (2012a,b), Chetty et al. (2014a), Kinsler (2016)), which will likely further increase their use in research and policy.

Because this paper takes as given that the data generating process is consistent with

[7]The recent outcry about a case where teacher value-added was incorrectly calculated in Washington DC, which resulted in firing mistakes (Strauss (2013)), evinces the considerable public concern about misclassifying public school teachers.

a value-added technology, it does not seek to contribute to the important, growing, and controversial (Rothstein (2017) and Chetty et al. (2017)) literature studying the statistical properties of value-added models. Rather, the contribution is to analyze the performance of widely used estimators of teacher quality within the value-added framework, which is the workhorse of existing teacher incentive schemes and education research. Therefore, the focus of this paper is quite different from, and complementary to, the literature studying whether value-added models are misspecified.[8] A different concern is that raised by Bond and Lang (2013), who question whether value-added 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, documented in, e.g., Fryer Jr. and Levitt (2004). In this paper, comparisons are made within one academic year (and within the same subject), mitigating this concern.[9]

Manski (2004) writes that "statisticians studying estimation have long made progress by restricting attention to tractable classes of estimators; for example, linear unbiased or asymptotic normal ones," (page 1231). In the same vein, I consider a set of estimators containing a (perhaps) naive estimator based on the "raw" data, which in a value-added framework would correspond to unbiased fixed effects, and the ubiquitous mean-square-minimizing empirical Bayes. I show below in Remark 1 that this is a natural set to consider.[10] The *unconstrained* class of optimal estimators would potentially condition on all available information, because the administrator could simply ignore information that was not valuable (Hölmstrom (1979)). No existing teacher incentive scheme does this. Therefore, in each environment, I focus on the less trivial and more relevant case where the administrator chooses the (constrained) optimal estimator from the set described above.

Teacher quality is distributed according to $\theta_i \sim F = \mathrm{N}(0, \sigma_\theta^2)$, where F is known.[11] As discussed in the introduction, the number of students assigned to teacher i, n_i, may depend on i's quality. For simplicity, I assume that class size depends on θ, where I sometimes denote this dependence by writing $n(\theta)$.[12] Note that what matters is the end relationship

[8] The maintained assumption that the technology is value-added also does not allow student ability to evolve flexibly; see Hansen et al. (2004), Todd and Wolpin (2007), and Ding and Lehrer (2014) for research relaxing this potentially important assumption.

[9] Moreover, comparisons are ordinal in the cutoff-based and hidden-type environments. However, it should be noted that the value-added estimates themselves are obtained using student-level data, which are assumed to have a cardinal structure.

[10] The framework developed here could also be used to study other classes of estimators.

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

[12] If class size were instead a noisy signal of teacher quality, the model solution would be more complicated without changing which estimator the administrator would prefer. I assume in this section that the administrator cannot directly condition on class size; Remark 2 discusses this assumption.

$n(\theta)$; whether it is the result of school principals assigning smaller class sizes to certain teachers or, say, teacher lobbying effort does not affect the results.

The test score gain for student j assigned to teacher i is $y_{ji} = \theta_i + \epsilon_{ji}$, where measurement error $\epsilon_{ji} \sim \mathrm{N}(0, \sigma_\epsilon^2)$ and $\epsilon_{ji} \perp\!\!\!\perp \theta_i$. I adopt this spare technology to simplify exposition; the quantitative results use value-added estimates that control for many characteristics.[13] I follow the literature and assume normal distributions due to two key properties that lead to considerable tractability: i) the normal distribution is closed under summation, which means each teacher's average measurement error will also be normally distributed and ii) the normal distribution is self-conjugate, which means the posterior distribution of (normally distributed) teacher quality, when it is added to a normally distributed average measurement error, is also normal. Results would likely also apply to other common symmetric, unimodal distributions.

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, given true quality θ_i, 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 value-added updates the prior (i.e., population) distribution of θ_i with data $\{y_{ji}\}_j$. Because both the prior distribution and measurement errors are normal, the posterior distribution is also normal, giving $\hat{\theta}_i^{EB} = \lambda_i \hat{\theta}_i^{FE} + (1 - \lambda_i)\,\mathrm{E}[\theta] = \lambda_i \hat{\theta}_i^{FE} = \lambda_i \cdot (\theta_i + \bar{\epsilon}_i)$, 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 (signal) relative to the estimated variation using the fixed effects estimator (signal plus noise).[14] I express the dependence of the weights on class size by writing $\lambda(n(\theta))$ or $\lambda(n_i)$, or the reduced-form $\lambda(\theta)$, depending on which is more convenient. How much the empirical Bayes 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 to be shifted to the sample mean. Note the empirical Bayes estimate for a particular teacher's quality is biased, i.e., $\mathrm{E}_{\bar{\epsilon}}\left[\hat{\theta}_i^{EB}\right] = \lambda(\theta)\theta_i \neq \theta_i$, but also has a lower variance. Though the exposition here is for fixed effects and empirical Bayes estimators, this bias-variance tradeoff would also apply to comparisons of other shrunken versus unshrunken estimators.

The fixed effects and empirical Bayes estimators differ only by the weights λ, making it simple to also consider estimators with intermediate weights by considering convex combinations of $\hat{\theta}_i^{EB}$ and $\hat{\theta}_i^{FE}$, resulting in a set of candidate estimators $[\hat{\theta}_i^{EB}, \hat{\theta}_i^{FE}]$. I obtain the optimal estimator by first analyzing the end points of this set (i.e., fixed effects and empirical Bayes) and then considering whether interior weights would be optimal.

[13] Students (and their parents) are assumed to be passive; in particular, they are assumed not to respond to changes in teacher quality. See Todd and Wolpin (2003) for a discussion of how this assumption could affect estimates from value-added models; see Ding and Lehrer (2010) for an application estimating dynamic treatment effects using Project STAR data.

[14] McCaffrey et al. (2003) discusses the differences between fixed effects and empirical Bayes estimators.

Remark 1 (Class size)**.** *Note that because the estimators only differ by λ_i, which in turn only differs between teachers via class size n_i, the analysis will focus on variation in class size without loss of generality.*

Remark 1 identifies what will be a common thread across the models in this paper: differences in estimator-specific value only depend on the relationship between class size and teacher quality. Intuitively, this relationship determines the amount of information in each estimator that is relevant to the administrator's objective, which may in turn affect her rankings over the estimators.

3 Cutoff-Based Model

In this section I develop a cutoff model to formalize the objective of a utility-maximizing decisionmaker, a school-district administrator, and characterize her optimal cutoff policy. I then use this model to evince the relationships between (i) how class size varies with teacher quality, (ii) her choice of estimator, and (iii) her expected maximized utility, i.e., value. The administrator takes as given an exogenous *desired cutoff* (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 may depend on estimator type, to maximize her expected objective over all teachers in the district.[15]

I begin with this model for several reasons. First, as will be shown below, her objective can be measured in terms of the number of correct and incorrect classifications with respect to the desired cutoff, embedding the administrator's objective in a natural metric: the expected number of mistakes. Second, a discrete policy is a natural fit for modeling discrete real-world policies like retention, making the analysis in this paper highly relevant for the most pervasive, and perhaps the most contentious, education policy debates.[16] Third, even though they are not obliged to take such a form, almost all existing teacher incentive schemes for public school teachers are cutoff-based, making this model's results immediately applicable to the vast majority of existing teacher incentive schemes; as noted by Stiglitz (1991) and Ferrall and Shearer (1999), real-world incentive schemes typically take very simple forms. Fourth, related literature also considers cutoff-based policies, e.g., Staiger and Rockoff (2010),

[15]Other work compares the statistical performance of different methods of estimating value-added. Schochet and Chiang (2012) calculate error rates for fixed effects and empirical Bayes estimators of teacher quality, assuming a fixed (identical) cutoff policy. Tate (2004) notes that ranks formed by fixed effects and empirical Bayes may differ depending on class size, but does not embed the analysis within a decision problem. Guarino et al. (2015) compare the performance of fixed effects and empirical Bayes estimators, with a focus on how they perform when students are not randomly assigned to teachers.

[16]Section 4.1.3 explores similarities between the cutoff-based model and the hidden type environment.

Hanushek (2011), Tincani (2012), Chetty et al. (2014b), and Rothstein (2014). Finally, the cutoff-based model's flexibility allows us to be agnostic about what underlies variation in measured output, which could be due to heterogeneity in fixed teacher productivity types and/or unobserved actions.

Model Specification The administrator receives utility from correctly rewarding a teacher with true quality equal to or higher than the *desired cutoff* κ (not making a Type I error) and not rewarding a teacher with true quality below κ (not making a Type II error). The administrator's utility from using estimator $\hat{\theta}$ and *cutoff policy* c on a teacher of true quality θ is:

$$u_{CP}(\theta, \hat{\theta}; c, \kappa) = \alpha \underbrace{1\{\hat{\theta} \geq c | \theta \geq \kappa\}}_{\text{avoid Type I error}} + (1-\alpha) \underbrace{1\{\hat{\theta} < c | \theta < \kappa\}}_{\text{avoid Type II error}},$$

where α and $(1-\alpha)$ are her weights on not making Type I and II errors, respectively. The parameter α helps link the model to the institutional context. An administrator tasked with firing the lowest-quality teachers might be willing to make many more Type I errors to avoid a Type II error (i.e., $\alpha < 1/2$). Alternatively, a high value of α may be more appropriate for an administrator allocating performance bonuses from a tight budget. If $\alpha = 1 - \alpha = 1/2$ the administrator values Type I and II errors equally.

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

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

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

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

For either estimator, an increase in the prospective cutoff policy c decreases the probabil-

ity 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 policy 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

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

The optimal cutoff for the fixed effects estimator c^{*FE} solves (3) when $\lambda(\theta) = 1, \forall\theta$. Denote the value to the administrator of using the optimal cutoff policies c^{*FE} and c^{*EB} as $v_{CP}^{FE}(\kappa) = \mathrm{E}\left[u_{CP}(\theta, \hat{\theta}^{FE}; c^{*FE}, \kappa)\right]$ and $v_{CP}^{EB}(\kappa) = \mathrm{E}\left[u_{CP}(\theta, \hat{\theta}^{EB}; c^{*EB}, \kappa)\right]$, 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, i.e., $v_{CP}^{FE}(\kappa) = v_{CP}^{EB}(\kappa) = 1$ for all desired cutoffs κ (Online Appendix B.2).

Remark 2 (Full information)**.** *This analysis assumes the administrator chooses a cutoff policy based on only test score information, e.g., she cannot directly condition on class size. The simplicity of such a policy makes it of obvious policy relevance, as is shown in Appendix Table A.1, which documents existing incentive pay programs and shows that none condition on class size. Additionally, when compared with a policy that may also explicitly condition on class size, a test-score-based cutoff could attenuate issues of class size manipulation for the sake of affecting the administrator's posterior about the quality of a particular teacher. However, because this assumption means empirical Bayes may be misspecified, in Online Appendix B.1 I consider a case in which the administrator can also directly condition on class size. Intuitively, the administrator would do no worse with this extra information, as she could always choose to ignore it. Because the obvious answer obtained in this scenario renders it of limited theoretical interest, the analysis in this paper focuses on estimators and policies that do not directly condition on class size.*[17]

[17] Just as the full-information MLE estimator of teacher quality would perform better than the shrinkage estimator, it may be the case that other statistical methods (such as supervised machine learning; see, e.g., Athey (2017)) could also improve upon the shrinkage estimator. Notably, predictions from either could depend on class size in a somewhat flexible form. This flexibility, however, comes at the cost of increasing the technical demands for implementation and, in the case of machine learning, using a "black-box" mapping from the data to the predicted value. Nevertheless, insofar as the machine learning techniques continued to "shrink" the raw data, the theoretical insights of this paper would likely still apply.

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 value is the same under both estimators. Next, I consider the case where class size depends on teacher quality. Proposition 2 shows that the administrator's relative value of the estimators depends on the relationship between class size and teacher quality, and also shows when the administrator would prefer either estimator for relevant ranges of parameters.

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

Proof. If all classes are the same size then $\lambda(n(\theta)) = \lambda, \forall\theta$. 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 (i.e., $v_{CP}^{FE}(\kappa) = v_{CP}^{EB}(\kappa)$). □

Note that Proposition 1 implies that the administrator would also be indifferent to using any convex combination of the estimators.

Proposition 2. *The administrator's preferred estimator in the cutoff model depends on the relationship between teacher quality and class size. In particular, when the relationship between teacher quality and class size is negative-(positive-)quadratic, the administrator will prefer the fixed effects (empirical Bayes) estimator for a wide range of parameter values.*

Proof. Because λ is strictly increasing in n, to simplify the proof's exposition I parameterize the empirical Bayes weights $\lambda(\cdot)$ directly as a function of θ, by assuming there is one slope for the relationship between teacher quality and weight below the population mean (β_-) and another slope for the relationship above the population mean (β_+), where either slope can be positive, negative, or zero; in Online Appendix B.6 I show that the results are not sensitive to this piecewise-linear specification. I also set $\sigma_{\bar{\epsilon}}(\theta) = \overline{\sigma}_{\bar{\epsilon}}$ for all teachers for the proof of the current proposition; in Online Appendix B.7 I show why this homoskedasticity assumption does not drive the result (regardless, $\sigma_{\bar{\epsilon}}$ varies between teachers in the numerical solution of the model and quantitative results). The empirical Bayes weight is then

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

where $\underline{\lambda} > 0$.

First, suppose that $\kappa < 0$ and that $c^{*EB} < 0$. Differentiating the administrator's value with respect to β_- and evaluating at $\beta_- = 0$ (which is a natural point at which to evaluate

the derivative because it takes as the starting point the constant class size case considered by Proposition 1), we obtain

$$\frac{\partial v_{CP}^{EB}}{\partial \beta_-}\Big|_{\beta_-=0} = (1-\alpha)\int_{-\infty}^{\kappa} \frac{-c^{*EB}\theta}{(\delta_-)^2\overline{\sigma_\epsilon}}\phi\left(\frac{\theta - \frac{c^{*EB}}{\delta_-}}{\overline{\sigma_\epsilon}}\right)\frac{\phi(\theta/\sigma_\theta)}{\sigma_\theta\Phi(\kappa/\sigma_\theta)}d\theta$$
$$+\alpha\int_{\kappa}^{0} \frac{c^{*EB}\theta}{(\delta_-)^2\overline{\sigma_\epsilon}}\phi\left(\frac{\theta - \frac{c^{*EB}}{\delta_-}}{\overline{\sigma_\epsilon}}\right)\frac{\phi(\theta/\sigma_\theta)}{\sigma_\theta(1-\Phi(\kappa/\sigma_\theta))}d\theta,$$

because $\frac{\partial v_{CP}^{EB}}{\partial c^{*EB}} \cdot \frac{\partial c^{*EB}}{\partial \beta_-} = 0$ due to the Envelope Theorem. The first term is negative because $-c^{*EB}\theta < 0$ for $\theta < \kappa$. Analogously, the second term is positive. If α is not too extreme, the first term will typically dominate, in which case the administrator's value is decreasing in β_-, i.e., the stronger is the increase in class size from teacher quality. Although the lack of a closed-form expression for the normal CDF precludes the construction of exact conditions on parameters under which the first term would dominate, Online Appendix B.3 provides sufficient conditions under which the first term would dominate, which apply to a wide range of underlying parameterizations of the cutoff model.

Analogously, if $\kappa > 0$ and $c^{*EB} > 0$ then by differentiating the administrator's value with respect to β_+ we can see that the administrator's value is increasing in β_+, meaning that increasing the weight associated with teacher fixed effects for teachers above the population mean improves the administrator's value. Online Appendix B.3 shows that the same conditions apply for this case as for when $\kappa < 0$.

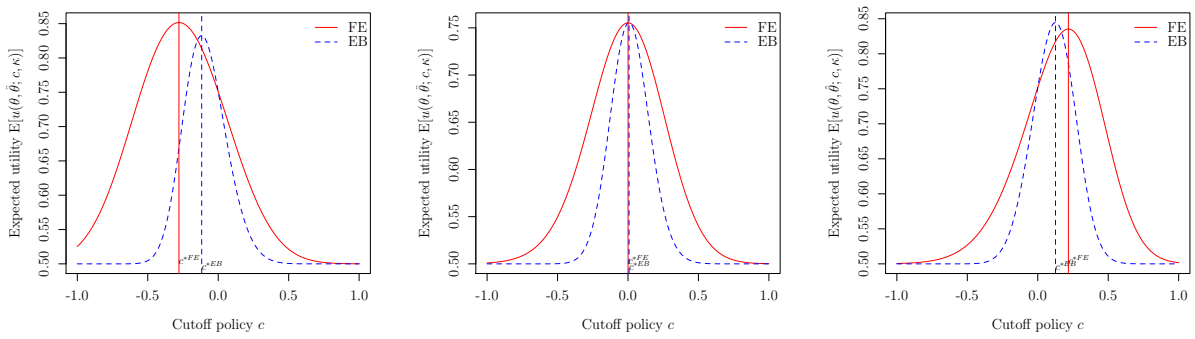
Therefore, reducing the slope of class size in teacher quality for below-average teachers and increasing the slope of class size in teacher quality for above-average teachers improves the administrator's utility from using the empirical Bayes estimator. In particular, starting from a constant class size, if we then shifted $\beta_- > 0$ and $\beta_+ < 0$ (shaped like a negative-quadratic relationship), the fixed effects estimator would provide the administrator with higher expected utility. On the other hand, if we shifted $\beta_- < 0$ and $\beta_+ > 0$ (shaped like a positive-quadratic relationship), she would prefer empirical Bayes. □

Note that interior convex combinations of the estimators will not be optimal. Intuitively, if one estimator is better at identifying certain teachers than the other, an intermediate estimator would also be outperformed by the corner.

Figure 1 illustrates Proposition 2 by plotting the expected utility of the administrator under the fixed effects estimator (solid red curve) and the empirical Bayes estimator (dotted blue curve) as a function of the cutoff policy for each estimator (x-axis), where class size is increasing in teacher quality, i.e., $\beta_-, \beta_+ > 0$, meaning that lower-quality teachers are weighted closer to the population mean than higher-quality teachers. Each curve traces

Figure 1: Administrator's objective, assuming class size increasing in teacher quality

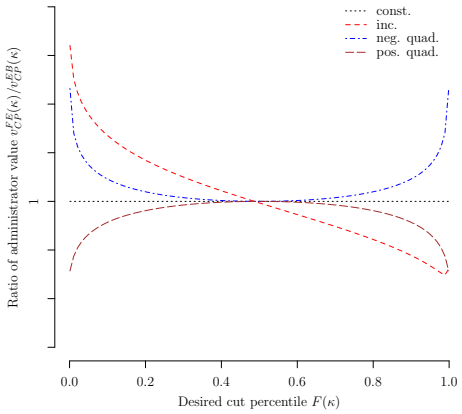
(a) Desired cut 1% (b) Desired cut 50% (c) Desired cut 95%



out the administrator's expected utility as a function of cutoff policies, given an exogenous desired cutoff quality κ. 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, $v_{CP}^{*\text{estimator}}(\kappa)$, is the maximum of each curve. If the administrator desires to separate the lowest quality teachers from the rest (Figure 1a), the re-weighting inherent in the empirical Bayes estimator can actually reverse teacher rankings and lead to a lower expected objective for the administrator than when the fixed effects estimator is used. The opposite is true for when the administrator wishes to separate the top teachers from the rest (Figure 1c); here, the peak of the empirical Bayes curve is higher. Intuitively, the empirical Bayes estimator is now dilating the estimated teacher quality further than the fixed effects 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 1b), fixed effects and empirical Bayes both obtain the same maximum height, i.e., they return the same expected objective. An increase in either δ_- or δ_+ corresponds to 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 2 summarizes the theoretical results for the cutoff model by comparing the performance of the estimators by plotting the ratio in value functions for the administrator $(v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa))$ 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 (average class size is the same across scenarios). For simplicity, α has been set to 1/2; Online Appendix B.4 shows this does not drive the findings for the vast majority of α values. Of course, if α took an extremely high value (i.e., $\alpha \to 1$), the second term in $\frac{\partial v_{CP}^{EB}}{\partial \beta_-}$ above would dominate; intuitively, if the administrator did

Figure 2: Ratio of administrator's value under fixed effects and empirical Bayes, by class size scenario and desired cut point



not value correctly identifying teachers below κ, their value would increase in β_-.[18]

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 $v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa)$ corresponding to the desired cutoff associated with the desired cut percentile $F(\kappa)$. As shown before, when class size is constant (dotted black line), the empirical Bayes cutoff is just a scaled version of the fixed effects cutoff and the administrator's value is the same under fixed effects and empirical Bayes estimators—i.e., $v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa) = 1$ for all κ. When class size is increasing in teacher quality (short-dashed red curve), the fixed effects estimator performs better than the empirical Bayes estimator when the administrator wishes to separate teachers of low quality from the rest (Figure 1a), while the empirical Bayes estimator performs better when the administrator wishes to isolate high-quality teachers (Figure 1c). When class size has a negative-quadratic relationship with teacher quality (dot-dashed blue curve), similar to the case in Proposition 2 where $\beta_- > 0$ and $\beta_+ < 0$, it is increasing when teacher quality is low and decreasing when teacher quality is high; in the example considered in Figure 2, the fixed effects estimator outperforms the empirical Bayes estimator at both the lowest and highest desired cutoffs. Finally, when class size is a positive-

[18] In their study of performance pay in education, Macartney et al. (2016) contrast policies focused on identifying immutable teacher "ability" and those providing sharp incentives for teachers to increase their provision of effort. The administrator's underlying policy goal (which, in the context of Macartney et al. (2016), would depend on the open question of whether unobserved teacher ability or unobserved effort was quantitatively more important) could be used to determine the appropriate (α, κ) in my framework, which then could be used to determine which would be the administrator's preferred estimator. That being said, Online Appendix B.4 documents the robustness of the administrator's preferred estimator as a function of class size scenario to even fairly asymmetric values for α (i.e., far from $\alpha = 1/2$) over a wide range of κ, which allows this paper a modicum of agnosticism regarding the administrator's exact policy goals.

quadratic function of teacher quality (long-dashed brown curve), the opposite is true. Figure 2 also demonstrates that the difference between the performance of fixed effects and empirical Bayes 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 fixed effects and empirical Bayes estimators when the administrator seeks to identify teachers as being on either side of the population mean (see Proposition 1 in Online Appendix B.5 for a proof that the administrator would be indifferent if her problem is *symmetric*).

Remark 3 (Pay for percentile)**.** *The cutoff model could be applied to a tournament-based scheme, e.g., "pay-for-percentile" (Barlevy and Neal (2012)), by considering an arbitrarily large sequence of desired cutoffs and associated bonuses for being above them. Therefore, the above results are also relevant for practitioners considering the design of such schemes or other, potentially continuous, ones. In such settings, the relevant quality distribution would be the one emerging in equilibrium, in response to the scheme.*

Remark 4 (One-period model)**.** *The model developed here is for one period. Although using an arbitrarily large number of periods when attempting to classify teachers would increase estimator precision and, hence, administrator value, doing so would preclude using schemes for many important decisions, such as termination of extremely low-quality inexperienced teachers. Note that although the model is for one period, with appropriate adjustments to its parameterization it could be applied to teachers as they progressed through their careers, which could accommodate, e.g., drift processes in the evolution of teacher quality (Wiswall (2013); Papay and Kraft (2015)).*

4 Asymmetric Information Models

Although the cutoff model considered in Section 3 has a close link to existing policy that is clearly desirable, it does not directly link measurement and output. Therefore, in this section I use the two main types of asymmetric information models to directly study how measurement of teacher quality may affect net output for a utility-maximizing administrator. Section 4.1 considers a hidden type model, in which unobserved types determine teacher quality. Section 4.2 considers a hidden action model, in which teachers take an unobserved action that determines their quality.

4.1 Hidden Type Model

This section develops a hidden type, or adverse selection, model.[19] It starts by considering a general version, Model HT-G, which derives the administrator's optimal policy when she can observe a fairly general output signal. In contrast to the cutoff model, where the administrator was assumed to follow a cutoff policy, this section shows that a cutoff-type policy would emerge as the optimal one in a general hidden type environment. This is useful because if a certain type of policy is optimal for the general signal in Model HT-G, then it would also be optimal for the specific estimators considered in subsequent sections.

4.1.1 Model HT-G

There are T periods, indexed by t, and J classrooms, or slots, indexed by j, where slot j has n_j students. As in the cutoff model, the administrator can provide rewards (or sanctions) to teachers, but class sizes may be determined by school principals. As in the real world, the administrator may condition on quality signals but not directly on other data, e.g., class sizes. Let I denote the set of potential teachers, or applicants, who are indexed by i. Per-student output from slot j being filled by teacher i in period t is $q_{it} = \beta_0 + \theta_{i(j,t)}$, where θ_i is teacher i's quality and output for slot j is zero if it has not been assigned a teacher (i.e., $i(j,t) = \emptyset$). The quality of applicants for teaching positions is distributed according to $\theta_i \sim \mathrm{N}(\mu, \sigma_\theta^2)$, where, as in the cutoff model, $\mu = 0$. Any teacher i in the applicant pool would accept a teaching job if offered a wage at least as high as $\underline{w}$ and there is an arbitrarily large number of teachers for each slot.

Teacher quality is not directly observed by the administrator, who, after the end of each period only observes a noisy signal of mean output $\hat{q}_{it} \sim G_{\hat{q}}(\hat{q}_{it}|q_{it})$. As in the cutoff model, the distribution of the output signal depends on true output q. However, I make a weaker assumption here, that $G_{\hat{q}}$ is a quasiconcave and symmetric error distribution.[20] Hiring a teacher costs χ output, where $\chi > 0$. Let I_t denote the subset of I who are employed as teachers in t. Let H_{it} denote the history of signals for teacher i that are observed at the beginning of period t, i.e., $H_{it} = \{\hat{q}_{i\tau}\}_{\tau<t}$, where the number of previous signals for i is $|H_{it}|$.

In each period, the administrator chooses a hiring policy $\psi_{h,t}(\cdot)$, where hiring occurs at the beginning of the period, and a reward policy $\psi_{r,t}(\cdot)$ to maximize her expected objective,

[19] This environment is similar to one developed in Staiger and Rockoff (2010).

[20] It might seem more intuitive to instead assume that $G_{\hat{q}}$ satisfied the Monotone Likelihood Ratio Property (MLRP). Karlin and Rubin (1956) note that many commonly used distributions satisfy the MLRP; examples include the normal, binomial, Poisson, and Gamma distributions, which explains this assumption's ubiquity in asymmetric information models (Chambers and Healy (2012)). However, under any of these distributions, $G_{\hat{q}}$ would only be guaranteed to satisfy the MLRP in the case of homoskedastic errors, which would preclude there being variation in class size.

where $\psi_{r,t}(\cdot)$ consists of a wage $w_{i(j,t)}$, paid at the beginning of the period, and a retention decision, made after that period's signals have been realized. The administrator chooses $\{\psi_{h,t}(\cdot), \psi_{r,t}(\cdot)\}_{t \in T}$ to maximize expected discounted total output, net the cost of her policy:

$$u_{HTG} = \sum_t \delta^{t-1} \mathrm{E}_t \left[\sum_j q_{i(j,t),t} - w_{i(j,t)} - 1\{|H_{i(j,t),t}| = 0\}\chi \right], \tag{4}$$

where δ is the discount rate, $\mathrm{E}_t[\cdot]$ denotes the expectation using information available at period t, and $|H_{i(j,t),t}| = 0$ means i is a new hire in period t.

Theoretical Results For simplicity, assume $\beta_0 = 0$ and set $\underline{w} = 0$.[21] Then, $\psi_{h,t}(\cdot)$ will be a list of $|J_t|$ random numbers for indices $i \in I/I_t$, where J_t denotes the set of empty slots at the beginning of period t (i.e., $J_t = J$ in the first period and then the slots with just-dismissed teachers thereafter). Now consider the administrator's choice of how to reward a given portfolio of teachers, $\psi_r(\cdot)$. In general, $\psi_r(\cdot)$ could depend on all signals (i.e., from the most recent and also earlier periods) of all currently employed teachers, and may have a complicated functional form. Proposition 3 greatly simplifies the solution.

Proposition 3. *The administrator's optimal policy $\psi_{r,t}(\cdot)$, for $i \in I_t$, will have the reservation value property consisting a stopping region and, if $G_{\hat{q}}$ is quasiconcave and symmetric, a continuation region above.*

Proof. First, note that the additive separability of (4) implies we can split it into J separate problems. Examination of (4) shows that the administrator's objective is increasing in output q_{it}, and therefore also increasing in expected output. If $G_{\hat{q}}$ is quasiconcave and symmetric, this implies that $\frac{\partial \mathrm{E}[q_{it}|\hat{q}_{it}]}{\partial \hat{q}_{it}} > 0$, i.e., the posterior mean of a teacher's quality is increasing in the signal $\hat{q}_{it}$.[22] Lippman and McCall (1976) prove that the optimal policy for each problem has a reservation value property (see also Shiryaev (2007)). Then, there will then be a signal region in which the administrator will retain the teacher (i.e., a continuation region) and below which she will pay χ to replace her (i.e., a stopping region). Finally, within the continuation region note that the administrator would not gain from paying additional wages per each slot, meaning that $\psi_{r,t}$ will feature a wage payment of $w_{\psi_{r,t}} = \underline{w} = 0$ and the

[21] This assumption is consistent with the administrator leaving no slots empty. An alternative would be to assume β_0 is such that the administrator would find it optimal to fill an empty slot j with a random hire from the pool of applicants, i.e., expected output is $\beta_0 + \mathrm{E}[\theta] = \beta_0 + \mu > \chi + \underline{w}$. This would encumber the notation without changing the result.

[22] This property is known as "updating in the direction of the signal" (UDS), and was derived by Chambers and Healy (2012).

retention decision will have a reservation value property.[23] □

The optimality of a reservation-value policy is typical of optimal stopping problems, of which the current model is an example, and suggests a link with the cutoff model from Section 3.[24] However, the administrator's objective (4) is quite general, which complicates obtaining theoretical results about how the administrator would prefer to measure teacher quality and relating results from the hidden type model to those from the cutoff-based model. Therefore, in Section 4.1.2 I study Model HT-0, a version of Model HT-G with two periods and constant class sizes. Model HT-1, in Online Appendix C.1, shows how a multi-period model, which allows teachers to become more productive as they gain experience, can be mapped into a series comprised of the second period of different HT-0 models. Model HT-2, in Online Appendix C.2, extends HT-0 to examine the case of variable class sizes. As with Model HT-0, a multi-period version of Model HT-2 could be related back to the second period of Model HT-2.

4.1.2 Model HT-0

There are two periods ($T = 2$) and teacher quality is fixed over time. Each slot j holds $n > 0$ students, which corresponds to the constant class size scenario for the cutoff-based model. Output per slot is noisily measured according to $\hat{q}_{jit} = q_{jit} + \bar{\epsilon}_{jit}$, where $\bar{\epsilon}_{jit} \sim \mathrm{N}\left(0, \sigma_\epsilon^2/n\right)$ and $\mathrm{E}\left[\bar{\epsilon}_{jit}|q_{jit}\right] = \mathrm{E}\left[\bar{\epsilon}_{jit}\right] = 0$. Let $\rho = \sigma_\theta^2/(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n})$ be the signal reliability, i.e., the amount of information about teacher quality in the output measure.

Theoretical Results As with Model HT-G, in the first period, the administrator hires at random from the pool of potential teachers. Therefore, I focus on the second period and suppress the period subscript t and discount rate δ. In the second period, she can choose to either retain or replace each teacher $i \in I_1$ based on information from the first period. Proposition 3 shows the optimal solution has a reservation value property. Our goal then is to characterize the marginal signal $\underline{q}$ in the distribution of first-period signals $\hat{q}$.

Per slot, the administrator's second-period objective from reservation value policy $\underline{q}$ on

[23] Variation in n_j would not affect the optimality of a reservation value policy, provided $n(\theta)$ was symmetric around the mean of teacher quality, as is the case in the positive- and negative-quadratic class size and teacher quality scenarios.

[24] Note that, although the administrator updates her prior beliefs using teacher quality estimates—in particular, she is Bayesian—she has a choice of estimators, one of which is itself "Bayesian" (shrinkage estimator) and one that is "frequentist" (fixed effects); she is not effectively "double-shrinking" the estimator.

signal $\hat{q}$ with a current teacher of quality q (which equals θ) is

$$\underbrace{1\{\hat{q} < \underline{q}\}\left(\mathrm{E}\left[q|\text{new hire}\right] - \chi\right)}_{\text{dismiss teacher; fill slot immediately}} + \underbrace{1\{\hat{q} \geq \underline{q}\}q}_{\text{retain teacher}} = 1\{\hat{q} < \underline{q}\}(\underbrace{\mathrm{E}\left[\theta|\text{new hire}\right]}_{=\mu=0} - \chi) + 1\{\hat{q} \geq \underline{q}\}\theta. \quad (5)$$

Taking expectations over output q and the signal $\hat{q}$, we can write the administrator's value of using estimator $\hat{q}$ with replacement cost χ as

$$v_{HT0}^{\hat{q}}(\chi) = \max_{\underline{q}} \Phi\left(\frac{\underline{q}}{\sigma_{\hat{q}}}\right)(-\chi) + \left(1 - \Phi\left(\frac{\underline{q}}{\sigma_{\hat{q}}}\right)\right)\mathrm{E}\left[\theta|\hat{q} \geq \underline{q}\right]. \quad (6)$$

By setting $\hat{q} = \hat{\theta}^{FE}$, the sample mean of each teacher's observed signals during the first period, we can then use (6) to write the administrator's value from using the fixed effects estimator:

$$v_{HT0}^{FE}(\chi) = \max_{\underline{q}^{FE}} \Phi\left(\frac{\underline{q}^{FE}}{\sigma_{\hat{\theta}^{FE}}}\right)(-\chi) + \left(1 - \Phi\left(\frac{\underline{q}^{FE}}{\sigma_{\hat{\theta}^{FE}}}\right)\right)\frac{\sigma_\theta^2}{\sigma_{\hat{\theta}^{FE}}}\frac{\phi(-\underline{q}^{FE}/\sigma_{\hat{\theta}^{FE}})}{\Phi(-\underline{q}^{FE}/\sigma_{\hat{\theta}^{FE}})}, \quad (7)$$

using the result for a truncated bivariate normal distribution, $\mathrm{E}\left[\theta|\hat{\theta}^{FE} \geq \underline{q}^{FE}\right] = \frac{\sigma_\theta^2}{\sigma_{\hat{\theta}^{FE}}}\frac{\phi(-\underline{q}^{FE}/\sigma_{\hat{\theta}^{FE}})}{\Phi(-\underline{q}^{FE}/\sigma_{\hat{\theta}^{FE}})}$ (see Greene (2003)).

We can characterize the marginal signal $\underline{q}^{*FE}$ by noting the administrator would be indifferent between replacing or retaining a teacher with that signal. The administrator's expected utility from replacing slot j's teacher is $\mathrm{E}\left[\theta\right] - \chi = -\chi$ and her expected utility from retaining j's teacher is $\mathrm{E}\left[\theta|\hat{\theta}^{FE}\right]$, which is equal to $(1-\rho)\mu + \rho\hat{\theta}^{FE} = \rho\hat{\theta}^{FE}$ by Bayes rule. The administrator will then replace teacher i if and only if

$$-\frac{\chi}{\rho} \equiv \underline{q}^{*FE} > \hat{\theta}_{i(j,1)}^{FE}. \quad (8)$$

To see this, first suppose that $\chi = 0$, in which case the marginal teacher is of average quality of the existing stock of teachers; since hiring in the first period is random from the pool of applicants this means any teacher with quality expected to be below the population average (μ) would be replaced; increasing χ would lower this threshold.

4.1.3 Relation Between Preferred Estimator in Cutoff and Hidden Type Models

This section shows how results from the cutoff-based model may also obtain in the hidden type environment. There are two main cases, corresponding to the class size scenarios covered by the propositions in Section 3. Given the results from the cutoff model, when characterizing the optimal estimator, I consider the corners of the set of estimators contained by the fixed

effects and empirical Bayes estimators.

Constant n When class sizes are constant, the administrator is indifferent between using either estimator. This is formalized in Proposition 4.

Proposition 4. *The administrator receives the same value from both estimators for any replacement cost* χ *when class size is constant.*

Proof. To obtain the administrator's value from using the empirical Bayes estimator $\hat{q} = \hat{\theta}^{EB} \equiv \lambda_{HT0}\hat{\theta}^{FE}$, where $\lambda_{HT0} \equiv \rho$, adapt (6) for the distribution of $\lambda_{HT0}\hat{\theta}$:

$$\begin{aligned} v_{HT0}^{EB}(\chi) &= \max_{\underline{q}^{EB}} \Phi\left(\frac{\underline{q}^{EB}}{\sigma_{\hat{\theta}^{EB}}}\right)(-\chi) + \left(1 - \Phi\left(\frac{\underline{q}^{EB}}{\sigma_{\hat{\theta}^{EB}}}\right)\right)\rho\frac{\sigma_\theta^2}{\sigma_{\hat{\theta}^{EB}}}\frac{\phi(-\underline{q}^{EB}/\sigma_{\hat{\theta}^{EB}})}{\Phi(-\underline{q}^{EB}/\sigma_{\hat{\theta}^{EB}})} \\ &= \max_{\underline{q}^{EB}} \Phi\left(\frac{\underline{q}^{EB}}{\rho\sigma_{\hat{\theta}^{FE}}}\right)(-\chi) + \left(1 - \Phi\left(\frac{\underline{q}^{EB}}{\rho\sigma_{\hat{\theta}^{FE}}}\right)\right)\rho\frac{\sigma_\theta^2}{\rho\sigma_{\hat{\theta}^{FE}}}\frac{\phi(-\underline{q}^{EB}/(\rho\sigma_{\hat{\theta}^{FE}}))}{\Phi(-\underline{q}^{EB}/(\rho\sigma_{\hat{\theta}^{FE}}))}, \end{aligned} \tag{9}$$

where the second line follows because $\sigma_{\hat{\theta}^{EB}} = \rho\sigma_{\hat{\theta}^{FE}}$. Then, if $\underline{q}^{*FE}$ solves (7) then $\underline{q}^{*EB} = \rho\underline{q}^{*FE}$ must solve (9) and, notably, return the same value for the administrator, i.e., $v_{HT0}^{FE}(\chi) = v_{HT0}^{EB}(\chi)$. □

Therefore, as with Proposition 1 for the cutoff model, in Model HT-0 the administrator would obtain the same value from using either estimator when class sizes are the same for all teachers. Note also that the optimal empirical Bayes reservation signal $\underline{q}^{*EB}$ is shrunk toward the population mean by exactly the same amount as was the optimal empirical Bayes cutoff policy, suggesting an equivalence in optimal policies in the cutoff-based model and HT-0. We can show this by setting $\kappa = -\chi$ and finding a Type I error weight α^{equiv} such that $c^{*FE}(\kappa = -\chi, \alpha^{equiv}) = \underline{q}^{*FE}(\chi)$. Then it will also be the case that $c^{*EB}(\kappa = -\chi, \alpha^{equiv}) = \underline{q}^{*EB}(\chi)$.

Model HT-1, in Online Appendix C.1, shows how the results for Model HT-0—in particular, its relation to the cutoff-based model—can be extended to allow for multiple periods and changes in teacher output over time, say, due to the accumulation of teaching experience. This is formalized in Proposition 5.

Proposition 5. *Model HT-1 can be mapped to Model HT-0.*

Proof. See Online Appendix C.1. □

Thus, the administrator would be indifferent in her choice of estimator for HT-0 or HT-1, i.e., when class size is constant.

Variable n Ideally, we would know that if an estimator would be preferred for every parameterization of the cutoff model, given a class size scenario, it would also be preferred for any hidden type environment for that class size scenario. Propositions 1, 4, and 5 show this is the case with constant class sizes. Model HT-2 extends Model HT-0 to allow for nonconstant class sizes. For brevity, this model is developed and analyzed in Online Appendix C.2.

The results from Model HT-2 are strikingly similar to those from the cutoff model: (i) the administrator's preferred estimator depends on $n(\theta)$ (same as in the cutoff model), (ii) the preferred estimator does not depend on the specific parameterization of HT-2 for a wide range of parameters, given $n(\theta)$ (same as in the cutoff model), and (iii) given $n(\theta)$, the administrator would prefer the same estimator in the cutoff model as she would in HT-2. In sum, I find that the preferred estimator in the cutoff model, which depends on the class size scenario $n(\theta)$, would typically also be preferred in model HT-2.

This similarity is intuitive. In the cutoff model, the administrator will have a higher value when there are fewer Type I errors, which in the hidden type model corresponds to fewer teachers of high true quality with quality measures below the reservation signal (i.e., replacement costs are lower). Likewise, the administrator in the cutoff model will also have a higher value when there are fewer Type II errors, which in the hidden type model corresponds to fewer teachers of low true quality with quality measures above the reservation signal (i.e., output will be higher).

Finally, note that one could model an increase in T by decreasing χ (from the two-period model), as replacing teachers would become relatively less costly when compared to the future gains in output. Then, the fact that the administrator would have the same preferred estimator for HT-2 suggests that she would also prefer the same estimator for multi-period versions of HT-2. It is important to note that, while Model HT-2 has two periods, a similar transformation to that done in Model HT-1 could be used to model multiple periods and potential changes in teacher output due to experience. If an estimator was preferred in each period, then it would also be preferred when calculating the discounted value of the administrator's dynamic objective.

4.2 Hidden Action Model

Many teacher incentive schemes are predicated on inducing higher effort levels from teachers. This section, therefore, presents the workhorse CARA-Normal model of moral hazard, as developed in Bolton and Dewatripont (2005), to illustrate the potential role choice of estimator may play in affecting output in a hidden action setting. The solution of this

model is the same as that in Hölmstrom and Milgrom (1987), which shows that the optimal contract features an end-of-period payment linear in measured output. This section builds on Mehta (2018), which calibrates this model to quantify the potential gains resulting from implementation of the optimal contract, using fixed effects. I begin by sketching the model here.

Model Specification There is one period. The administrator has utility $q - w$, where q is output and w is the wage paid to the teacher. The teacher has constant absolute risk aversion (CARA) utility $-e^{-\xi(w-\psi(a))}$, where ξ is their coefficient of absolute risk-aversion and the cost of exerting effort a is $\psi(a) = \gamma a^2/2$. The teacher requires an expected utility of $\underline{u}$ to participate. Output from teacher i depends on teacher quality according to $q_i = \theta_i$, where teacher quality $\theta_i = a_i + \nu_i$. The term a_i is the teacher's endogenous effort level and the error $\nu_i \sim \mathrm{N}(0, \sigma_\nu^2)$ is a productivity shock common to students taught by the teacher; ν could correspond to a teacher-classroom-specific match effect. Assume ν can be observed by the school principal, meaning there may be a relationship between teacher quality and class size, as in the other models. The teacher chooses a, without knowing the realization of ν. Average output for teacher i is noisily measured according to an average test score $\hat{q}_i = q_i + \bar{\epsilon}_i = \theta_i + \bar{\epsilon}_i = a_i + \nu_i + \bar{\epsilon}_i$. Note that the risk-neutrality of the administrator's objective implies that she can solve a separate problem for each teacher.

Hölmstrom and Milgrom (1987) show that it is optimal for the administrator to pay the teacher based on the noisy output measure using a linear contract $w = \beta_0 + \beta_1 \hat{q}$, where β_1 is the share of measured output paid to the teacher. Note that, from the teacher's perspective, uncertainty comes from the composite error $\nu_i + \bar{\epsilon}_i$, which are collected as η_i. We can then write the wage as $w(a, \eta)$, where the administrator can only observe $a + \eta$. Ex-ante, teachers face the same uncertainty about η_i.[25]

Substituting for output and output measure and using the result that the optimal contract

[25] This section adopts the simplifying assumption that teachers treat η_i as being normally distributed when solving for their optimal action. Technically, they should integrate over the *distribution* of distributions of $\bar{\epsilon}_i$ if $n(\theta)$ is not constant. Simulation results confirm that η_i is approximately normally distributed for reasonable parameter values; a Kolmogorov-Smirnov test of normality of η_i has a p-value of 0.131. Further note that all teachers would still have the same equilibrium action in the latter case, meaning this assumption would not affect the qualitative predictions from this model.

will be linear in observed output, the administrator's problem is

$$\begin{aligned} \max_{\beta_0,\beta_1} \ & \mathrm{E}_{\nu,\eta}\left[a+\nu-w(a,\eta)\right] && (10)\\ \text{s.t.} \ & w(a,\eta)=\beta_0+\beta_1(a+\eta) \\ & \mathrm{E}_{\eta}\left[-e^{-\xi(w(a,\eta)-\psi(a))}\right]\geq \underline{u} && \text{(IR)}\\ & a\in \arg\max \ \mathrm{E}_{\eta}\left[-e^{-\xi(w(a,\eta)-\psi(a))}\right], && \text{(IC)} \end{aligned}$$

where the individual rationality constraint (IR) ensures participation and the incentive compatibility constraint (IC) characterizes the teacher's choice of action.

The teacher problem yields a unique optimal action $a^* = \beta_1/\gamma$ by differentiating (IC) with respect to action and the optimal linear contract features $\beta_1^* = 1/(1+\xi\gamma\sigma_\eta^2)$ (see pp. 137-139 of Bolton and Dewatripont (2005) for details).[26] Therefore, expected output is $\mathrm{E}[q^*] = \mathrm{E}_\nu[a^*+\nu] = a^* = 1/\left(\gamma(1+\xi\gamma\sigma_\eta^2)\right)$.[27] Intuitively, as the signal quality worsens (i.e., σ_η^2 increases) the contract becomes lower powered (i.e., β_1^* decreases), resulting in lower action a^* and expected output $\mathrm{E}[q^*]$.

As with the hidden type model, the choice of estimator may affect output in the hidden action environment. The fixed effects estimator would simply be the (unadulterated) output signal, i.e., $\hat{q}_i^{FE} = \hat{q}_i$. Proposition 6 considers the case of constant class sizes.

Proposition 6. *The administrator receives the same value from both estimators in Model HA when class size is constant.*

Proof. The empirical Bayes estimator would be $\hat{q}_i^{FE}$ shrunk by a constant factor λ, i.e., $\hat{q}_i^{EB} = \lambda\hat{q}_i$. If $(\beta_0^{*FE}, \beta_1^{*FE})$ solves (10) when using output measure $\hat{q}_i^{FE}$ then it must be that $(\beta_0^{*FE}, \beta_1^{*FE}/\lambda)$ solves (10) when using output measure $\lambda\hat{q}_i$. Thus, the administrator obtains the same value from using either estimator. □

Intuitively, empirical Bayes contains the same amount of information as fixed effects when class sizes are constant, meaning the contract slope would simply adjust to take into account its shrunken distribution. Model HA highlights the bias-variance "tradeoff" that has potentially been the source of confusion, leading to the adoption of shrinkage estimators in many applications. If the variance of the fixed effects estimator increased, the resulting

[26] Note that, according to this model, output will necessarily be zero when teachers are salaried (i.e., $\beta_1 = 0$), which is the case in many real-world applications in which, for various reasons, output-based pay has not been implemented. This obviously counterfactual implication can be resolved by assuming there are two types of effort: the action a which is only imperfectly measured and another action that is perfectly observed, and therefore, contractible.

[27] Note that, although in this moral hazard setting there is a degenerate distribution of teacher *effort* in equilibrium, measured teacher *quality* (i.e., average test score $\hat{q}$) is normally distributed.

optimal contract would partially protect a risk-averse teacher by making incentives weaker in the output measure (i.e., test scores), or reducing the slope of the linear contract β_1. The more risk-averse the teacher, the more protected they would be (i.e., the shallower the slope β_1). Crucially, the optimal contract would not respond to an increase in noise by "changing the data" (e.g., switching to a lower-variance estimator), but rather, would in equilibrium adjust the way in which the data were used in remuneration (i.e., decrease β_1).

The fact that Proposition 6 shows we can re-scale the empirical Bayes estimator when class size is constant suggests the use of a biased, yet lower-variance estimator could be modeled by increasing the effective error variance σ_η^2.

Proposition 7. *The administrator's preferred estimator in Model HA depends on the relationship between teacher quality and class size. In particular, when the relationship between teacher quality and class size is negative-(positive-)quadratic, the administrator will prefer the fixed effects (empirical Bayes) estimator.*

Proof. We can apply the informativeness principal of Hölmstrom (1979), which relates the value of a signal to how much information it contains, and rank estimators based on how much information they contain about teacher quality. I do this by examining the signal-to-noise ratio in the output measure. The empirical Bayes signal is $\hat{q}_i^{EB} = \lambda_i \hat{q}_i = \lambda_i \theta_i + \lambda_i \bar{\epsilon}_i$, which has a total amount of signal about θ (i.e., fraction of variation explained by θ) of

$$\int_{-\infty}^{\infty} \frac{[\lambda(\theta)\theta]^2}{[\lambda(\theta)\theta]^2 + [\lambda(\theta)\sigma_{\bar{\epsilon}}(n(\theta))]^2} dF(\theta). \tag{11}$$

The numerator of expression (11), i.e., the amount of signal contained in the measure, is smaller when $n(\theta)$ is negative quadratic and larger when $n(\theta)$ is positive quadratic.[28] □

Therefore, as with the cutoff and hidden type models, the theoretical effect of switching from empirical Bayes to fixed effects is clear in the hidden action model, given the relationship between class size and teacher quality: the administrator's value would be the same with constant class sizes, lower under empirical Bayes with a negative-quadratic $n(\theta)$, and higher under empirical Bayes when $n(\theta)$ is positive quadratic.

[28] As with the cutoff model, to simplify exposition I parameterize the empirical Bayes weights according to $\lambda(\theta) = \max\{\underline{\lambda}, \delta_- + \beta_- \theta\}$ if $\theta < 0$ and $\lambda(\theta) = \max\{\underline{\lambda}, \delta_+ + \beta_+ \theta\}$ if $\theta \geq 0$, where $\underline{\lambda} > 0$. In this case, the derivative of the numerator of the integrand with respect to β_-, evaluated at $\beta_- = 0$, is $2\theta^3\delta_- < 0$ and the derivative of the numerator of the integrand with respect to β_+, evaluated at $\beta_+ = 0$, is $2\theta^3\delta_+ > 0$.

5 Quantitative Results

In this section, I quantify the estimators' performance, using data from the Los Angeles Unified School District, the second-largest school district in the US.[29] In Section 5.1, I calibrate parameters needed to compare estimator performance in the cutoff model. In Section 5.2, 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. Section 5.3 presents a calculation of how choice of estimator would affect output in the hidden type model. Section 5.4 discusses calibration of the additional parameters of the hidden action model and computes how choice of estimator would affect output there. Although these incentive schemes are not currently in place in Los Angeles, these exercises can serve as a useful benchmark for how the estimators might perform when used in similar incentive schemes. Indeed, the fact that a high-stakes scheme was not in place obviates addressing the potential strategic re-assignment of students to teachers.

5.1 Calibration

The cutoff model shows that the difference in the administrator's value depends on the variances of teacher quality σ_θ^2 and the test score measurement error σ_ϵ^2 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.

Variances Schochet and Chiang (2012) compile estimates of the variances from a large number of studies in their study of error rates in value-added models, providing a good source for typical values for σ_θ^2 and σ_ϵ^2 (see Online Appendix D.1). The chosen parameter values of $\sigma_\theta^2 = 0.046$ and $\sigma_\epsilon^2 = 0.953$ indicate that the variance of the measurement error is about 20 times the size of the variance of teacher quality, resulting in an average student-achievement signal-to-noise ratio of 0.512. That is, student achievement for the average teacher in Los Angeles is about equal parts signal and noise. As has been noted by many other researchers studying a wide variety of contexts (e.g., McCaffrey et al. (2009), Staiger and Rockoff (2010)), it is difficult to correctly classify teachers.

Relationship Between Class Size and Teacher Quality I recover the relationship between class size and teacher quality using value-added estimates provided by the Los Angeles Times. In 2011, the Los Angeles Times published the results of a RAND Corporation study estimating value-added for over 30,000 teachers serving almost 700,000 students (Buddin

[29] Imberman and Lovenheim (2016) use these data in their study of the market's valuation of value-added.

(2011)).[30] The dataset contains estimated value-added, estimating using fixed-effects models, for 3rd to 5th grade teachers in both Reading and Math and class sizes which condition on several variables, including past performance of students, class size, student characteristics such as race, gender, English proficiency and parents' education, and classroom composition (past performance of classmates and their student characteristics as well).[31] In addition to describing the relationship between teacher quality and class size, which is critical to compare the performance of the estimators, the distributions of value-added estimates from Buddin (2011) are similar to those in Schochet and Chiang (2012).[32] The average class size is 22.5 students, with a standard deviation of 5 students.

Remark 5. *As is common in studies using a value-added framework, estimated teacher quality does not account for potential ceiling effects, which could affect value-added estimates if many students score at the upper bound of the test instrument. However, around 40-60% of California students score as proficient or higher on the test instrument, well below the region identified as problematic by Koedel and Betts (2010).*[33] *Therefore, ceiling effects do not likely play a large role in this paper's empirical application.*

Figure 3 plots non-parametric regressions (solid blue lines) of class size on estimated teacher value-added for Reading (3a) and Math (3b). Teachers at either end of the distribution of Reading value-added have the smallest class sizes and those in the middle of the distribution have the largest class sizes. Table 1 shows the results of regressions of teacher class size on estimated teacher quality and estimated teacher quality squared. The first two columns are for Reading and the second two are for Math. The dotted black lines on Figure 3 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, showing that the estimates from the full sample are not driven by outliers. These results indicate that class size is indeed increasing in value-added in the lowest part of the distribution and decreasing in value-added in the highest part of the distribution. The relationship is not as clear for math value-added, but the regression shows that class size first increases and then decreases for reading value-added, with a negative quadratic term for math value-added. Strikingly,

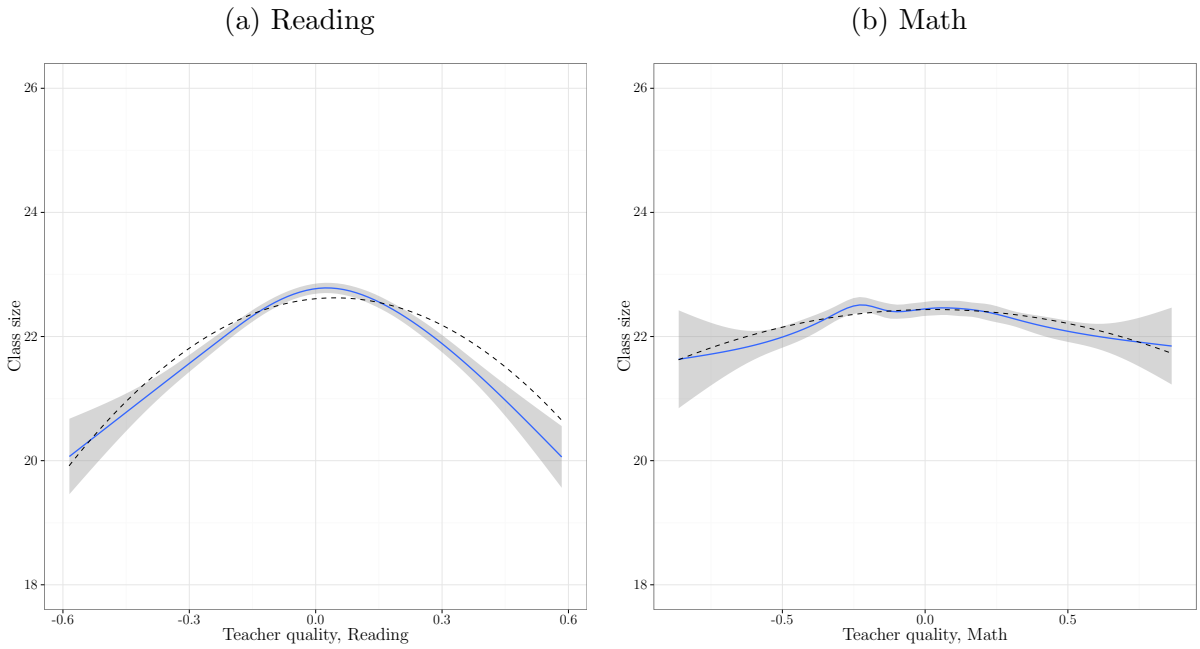
[30] `http://projects.latimes.com/value-added/`

[31] The results do not appreciably change when using value-added estimates from specifications that control for subsets of these characteristics. All specifications control for class size, which may result in an attenuated calibrated relationship between teacher quality and class size.

[32] The distributions of value-added in the data have means of 6.4E-11 and 1.3E-10 and variances of 0.038 and 0.083 for Reading and Math value-added, respectively. Because the quantitative results combine data from Buddin (2011) 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.

[33] See `https://www2.ed.gov/admins/lead/account/consolidated/index.html` for details.

Figure 3: The relationship between class size and teacher quality



Note: The solid blue lines plot predicted class size from non-parametric regressions of class size on estimated teacher value-added for Reading (left panel) and Math (right panel), with 95% confidence intervals shaded gray. The dotted black lines in each panel plot predicted class size from quadratic regressions of class size on estimated teacher value-added for the full sample, the results of which are presented in columns (1) and (3) of Table 1, for Reading and Math, respectively.

the observed relationship between teacher quality and class size is the worst-case scenario for the empirical Bayes estimator for the cutoff-based and asymmetric information models.

To most closely match the model, $n(\theta)$ would ideally be known and fed into the administrator's problem. In practice, only estimates of $n(\theta)$, denoted by $\hat{n}(\hat{\theta})$, are directly available from any dataset; the latter are what was presented in Figure 3 and Table 1. The estimated relationship $\hat{n}(\hat{\theta})$ also features a mechanical negative-quadratic relationship, caused by heteroskedastic errors that would be generated even under independently and identically distributed class sizes. Intuitively, even with class size n_i being distributed independently from teacher quality θ_i, teachers with very small class sizes would end up in the tails of the measured quality distribution because their mean errors would have higher variances. To address these issues, I calibrate $n(\theta)$ using an indirect inference approach described in Online Appendix D.2; this calibrated relationship is then used in the quantitative results presented later in this section. Table 2 presents the calibrated relationships between teacher quality and class size, $n(\theta)$. The first column presents the intercept, the second the linear term, and the third the term on the quadratic variable. The negative-quadratic term in the

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)

Note: ***p<0.01. Columns (1) and (3) present regression coefficients using the full samples of Reading and Math value-added estimates, respectively. Columns (2) and (4) present regression coefficients obtained when excluding teachers with estimated value added more than two standard deviations from the mean, for Reading and Math, respectively.

calibrated relationship between class size and teacher quality for Reading is stronger than that presented in Table 1, at -13.929, compared to -6.801 in column (1) of Table 1. On the other hand, there is a negligible relationship between class size and teacher quality in Math. That is, the mechanical relationship generated by heteroskedasticity can basically explain the fairly weak pattern in Table 1.[34]

5.2 Quantitative Findings: Cutoff Model

This section computes the administrator's value from using each estimator for a wide range of desired cutoffs, using the calibrated values of error variances and the relationship between class size and teacher quality obtained in Section 5.1. For each desired cutoff κ and subject (e.g., identifying teachers with quality at or above the 99th percentile for Reading value-

[34] Table 1 shows that both the Reading and Math negative-quadratic terms become larger in absolute value when excluding teachers with estimated quality two standard deviations and more from mean quality. Because the calibrated relationships are based on the full sample, they could therefore be viewed as conservative characterizations of the strength of the negative-quadratic relationship between teacher quality and class size.

Table 2: Calibrated $n(\theta)$, by subject

Subject	Constant	Subject quality	Sq. subject quality	Res. Std. Error
Reading	22.702	1.031	-13.929	5.124
Math	22.263	-0.225	-0.039	4.388

Note: Each row presents the calibrated coefficients for an equation expressing class size in a subject as a function of a constant, teacher quality in that subject, and teacher quality in that subject, squared. Calibration details are in Online Appendix D.2.

added), I solve for the administrator's optimal cutoff policy for fixed-effects and empirical Bayes estimators, assuming a symmetric loss function.[35] This returns an expected objective for each estimator, for each desired cutoff (and subject), i.e., $v_{CP}^{FE}(\kappa)$ and $v_{CP}^{EB}(\kappa)$ for the fixed-effects and empirical Bayes estimators, respectively (for Reading).

Figure 4a plots the ratio of the administrator's maximized expected objective under the fixed effects and empirical Bayes estimators ($v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa)$) for Reading (solid black curve) and Math (dotted red curve), for desired cutoffs ranging from the lowest to the highest teacher qualities. The right panel (4b) plots how many more expected mistakes (i.e., the expected sum of Type I and II errors) the empirical Bayes estimator would make than the fixed effects estimator, assuming the Los Angeles school district employed 30,000 teachers.[36] We can see that the quadratic nature of the association between teacher quality and class size affects the relative performance of the fixed effects and empirical Bayes 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 value of using fixed effects rather than empirical Bayes estimators. The administrator's value is higher almost everywhere when she uses the fixed effects estimator, and the relative performance of the empirical Bayes estimator is the worst at the extremes of the distribution of teacher quality. For example, the empirical Bayes estimator would make almost 800 more mistakes than fixed effects when the desired cutoff is at the 1st percentile, and 600 more when the desired cutoff is at the 99th percentile. Put another way, even when the administrator is allowed to re-optimize and choose an estimator-specific cutoff policy, using empirical Bayes would result in 9.5% more classification mistakes when the desired cutoff was at the 1st percentile of teacher quality and 7.3% more mistakes when the desired cutoff was the 99th percentile of teacher quality.[37] The administrator's values from using the fixed effects and

[35]Results are qualitatively similar under a wide range of asymmetric preferences, i.e., where $\alpha \neq 1-\alpha$; see Online Appendix B.4.

[36]Though the value-added 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`).

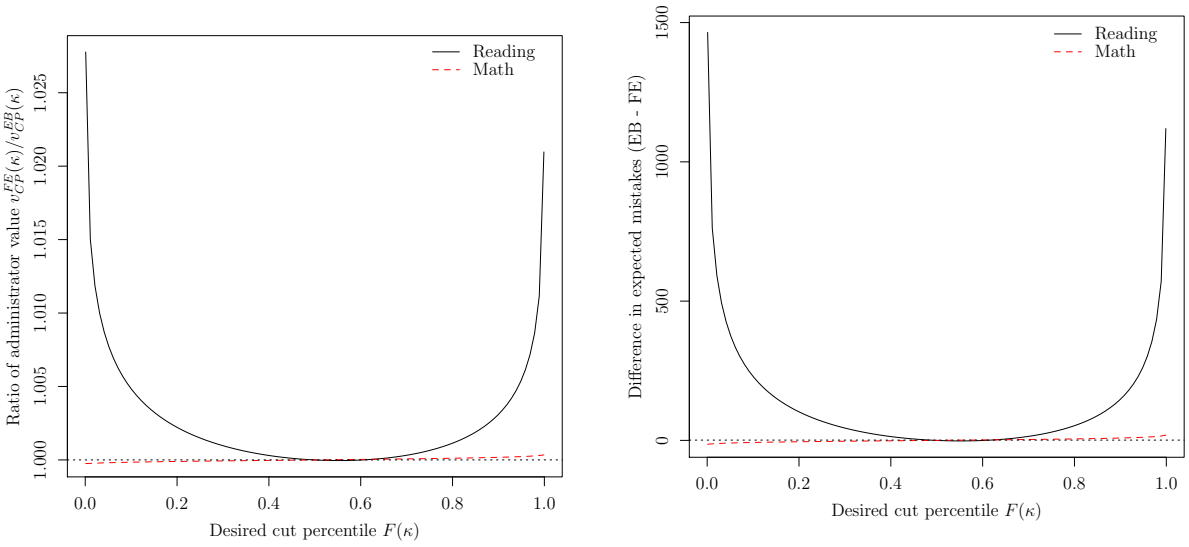
[37]The fraction of classification mistakes when using fixed effects when the desired cutoff κ is the 1st and 99th percentile would be 27.8% and 27.1%, respectively.

empirical Bayes estimators become comparable as the desired cutoff approaches the center of the distribution of teacher quality. The performance of the fixed effects and empirical Bayes estimators most greatly diverges precisely where policies that sanction very low-performing teachers or reward very high-performing teachers would bite the most, and the fixed effects estimator returns higher expected maximized utility (i.e., in expectation would make fewer mistakes) under almost every desired cutoff.

Figure 4: Administrator's value and difference in mistakes, using calibrated $n(\theta)$

(a) Ratio of administrator value $(v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa))$

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



Note: Number of decisions is 30,000.

Although the divergence in estimator performance is largest when the desired cutoff is in the tails of true teacher quality, all teachers would be affected by the administrator's choice of estimator. Figure 5a plots the probability that a teacher with true quality θ, measured along the x-axis, has an estimated quality $\hat{\theta}$ above the optimal cutoff policy corresponding to a desired cutoff κ of the first percentile of true teacher quality (dotted black line), e.g., $\Pr\{\hat{\theta}^{FE} \geq c^{*FE}\}$ for the fixed effects estimator. This desired cutoff could correspond to firing teachers with quality at or below the first percentile. These probabilities are plotted for the fixed effects (solid red curve) and empirical Bayes (dashed blue curve) estimators, using the relationship between class size and teacher quality for Reading. The shaded area corresponds to teachers with true quality below the desired cutoff. Having an estimated quality above c^* for teachers in this region would mean the administrator made a Type II error, e.g., they were incorrectly retained, the probability of which corresponds to the distance from the

estimator-specific curve to 1 in Figure 5a. For teachers outside the shaded region, having an estimated quality below c^* would correspond to a Type I error, e.g., they were incorrectly dismissed, the probability of which corresponds to the height of the estimator-specific curve.

For each estimator, the probability of having estimated quality above the optimal cutoff policy increases as a teacher's true quality increases (i.e., we move to the right). However, the fixed effects estimator has a higher probability of measuring above-threshold teachers as above c^{*FE} than does empirical Bayes for its corresponding optimal cutoff policy and a lower probability of measuring below-threshold teachers as above c^{*FE}. That is, fixed effects would have lower probabilities of both Type I and Type II errors. This is more clear in Figure 5b, which plots the ratio of probability of the estimate being above the respective cutoff for fixed effects over empirical Bayes, i.e., $\Pr\{\hat{\theta}^{FE} \geq c^{*FE}\}/\Pr\{\hat{\theta}^{EB} \geq c^{*EB}\}$.[38] For example, fixed effects would have a 40% lower chance of measuring a teacher with true quality more than four standard deviations below the mean ($\theta \approx -0.8$)—well below the desired cutoff quality of the first percentile—as above the optimal cutoff policy and a 10% higher chance of finding a teacher with true quality about 1.5 sd below the mean ($\theta \approx -0.3$)—above the desired cutoff quality—as above the cutoff policy.

5.3 Quantitative Findings: Hidden Type Model

This section (roughly) examines how the choice of estimator would affect net output in a hidden action environment, which requires information about the replacement cost χ. I computed the administrator's value under Model HT-2 (i.e., HT-0 with nonconstant $n(\theta)$) under fixed effects and empirical Bayes estimators using the calibrated Reading class size relationship from Table 2 and a calibrated replacement cost value of $\chi = 0.25\sigma_\theta = 0.054$. I chose this value for χ because Wiswall (2013) reports that teachers with 30 years of experience have value-added that is one standard deviation higher than new teachers and 0.75 standard deviations higher than teachers with five years of experience, implying a 0.25 standard deviation difference acquired in the first five years of experience. This value is similar to that used in Staiger and Rockoff (2010), who assume a first-year teacher has an average value-added 0.07 sd lower than teachers with two or more years of experience. Note that by setting χ in terms of standard deviations of teacher quality, the outcome is naturally viewed in terms of teacher quality, which has been shown to appreciably affect economic output (Hanushek (2011)).

[38] The crossing point of the curves in Figure 5a does not occur at the desired cutoff κ because, at the recovered $n(\theta)$ relationship, the empirical Bayes cutoff policy c^{*EB} is closer to zero than the fixed effects cutoff policy c^{*FE}, which means that teachers with qualities θ just below κ would be less likely to be measured as being above the empirical Bayes cutoff than the fixed effects cutoff.

Figure 5: Probability of being measured above optimal cutoff policy, given $F^{-1}(\kappa) = 0.01$

(a) Probability of being above c^*

(b) Ratio of probability of being above c^*, FE/EB

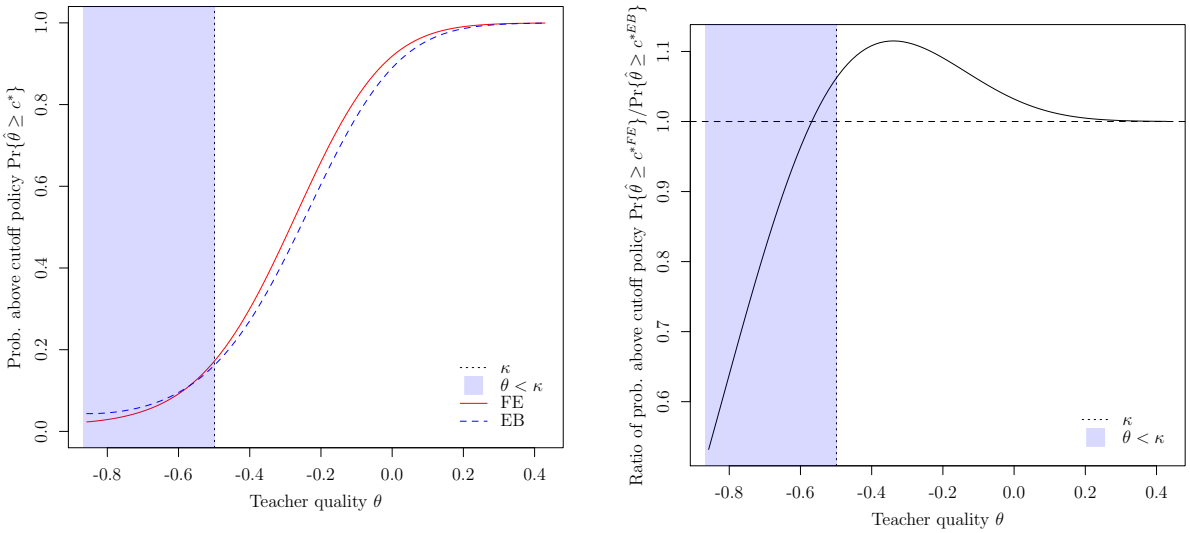


Figure 5a plots the probability that a teacher with true quality θ will be classified as being above the optimal estimator-specific cutoff policy c^* for a desired cutoff κ equal to the first percentile of teacher quality (the vertical, dotted black line). The solid red curve depicts the probability under fixed effects and the dashed blue curve depicts the probability under empirical Bayes. The shaded region indicates teachers with true quality below the desired cutoff. Figure 5b plots the ratios of the probabilities depicted in Figure 5a.

The administrator's value when using empirical Bayes and the optimal reservation signal policy $\underline{q}^{*EB}(\chi, n(\theta))$, is 0.038 of a standard deviation; teachers with quality measures in the bottom 37% would be replaced. The administrator's value from using fixed effects would be 0.5% larger, where this difference is driven by the lower optimal cost incurred under the fixed effects estimator, which corresponds to fewer classification mistakes at the low end of teacher quality. If instead, we used the value $\chi = 0.07$ from Staiger and Rockoff (2010), the reservation signal would decrease in response; here, teachers with the lowest 33% signals would be replaced. The administrator's value under empirical Bayes would be 0.032 standard deviations. Naturally, the larger replacement cost lowers the administrator's optimized objective. The administrator's value from using fixed effects would be 0.8% larger, where this difference is again driven by the lower cost of the optimal policy under fixed effects.

5.4 Quantitative Findings: Hidden Action Model

This section takes two approaches to (roughly) examine how choice of estimator might affect optimal output in a hidden action environment. First, it uses estimates from Muralidharan

and Sundararaman (2011) to calibrate parameters from the hidden action model. Second, it computes the effect on output from using either estimator of teacher quality for a wide range of model primitives. The approaches use the relationship between class size and teacher quality for Reading, from Section 5.1, and yield similar findings regarding the increase in output coming from the administrator's use of fixed effects, instead of empirical Bayes. Note that, in each approach, actions and output are measured relative to their baseline level, i.e., that provided by teachers absent output-based incentives.

In the hidden action model, output is a function of the action, which itself depends on the variance of noise η, CARA parameter ξ, and cost parameter γ. I first characterize how much information the administrator can extract about teacher quality (here, teacher effort choices) using either estimator. I do this by calibrating the implied variance of the composite error η for the fixed effects and empirical Bayes estimators (details are in Online Appendix D.3).[39] Based on Proposition 7, I model the information loss when using empirical Bayes under a negative-quadratic relationship between class size and teacher quality by increasing the measurement error variance on teacher action, σ^2_η, by 3.2%.[40]

Mehta (2018) uses data from Muralidharan and Sundararaman (2011), which estimates the effect of a linear output-based incentive scheme for teachers in the Indian state of Andhra Pradesh, and other information to calibrate the model parameters $(\gamma, \xi, \sigma^2_\eta)$. These are then used to characterize the optimal contract when using the fixed effects estimator and the gains from implementing the optimal contract, which at the calibrated parameters would be over six times larger than those in Muralidharan and Sundararaman (2011). Here, I take that calibration as given and compute the effect of using the empirical Bayes estimator on equilibrium output under the optimal contract.

Briefly, Mehta (2018) exploits the teacher's optimal choice of action, which solves (IC) in (10) but does not rely on optimality of the slope β_1, to map (β_1, a) to the cost $\gamma = 2.577 \times 10^{-5}$. The CARA parameter is set to $\xi = 6.7 \times 10^{-3}$, the mean estimated CARA from the benchmark model of Cohen and Einav (2007), Table 5. Finally, the variance of output is calibrated to σ^2_η =6,971,331\$^{2}. At the calibrated parameters, the optimal slope is $\beta_1^{*FE} = 0.454$, which has a corresponding optimal action of $a^{*FE} = \$17,608.83$; this corresponds to an average increase in student achievement of 1.063 sd. Either increase is more than six times larger than the estimated increase in student achievement stemming from the much weaker incentives provided under the experiment.

[39] This exercise abstracts from the error introduced by class size uncertainty, which would understate the gain in output from using fixed effects instead of empirical Bayes.

[40] Although it would be in principle possible to also directly condition on class size, as has been discussed previously, this would introduce a direct incentive to manipulate class size to affect the administrator's posterior beliefs about teacher quality.

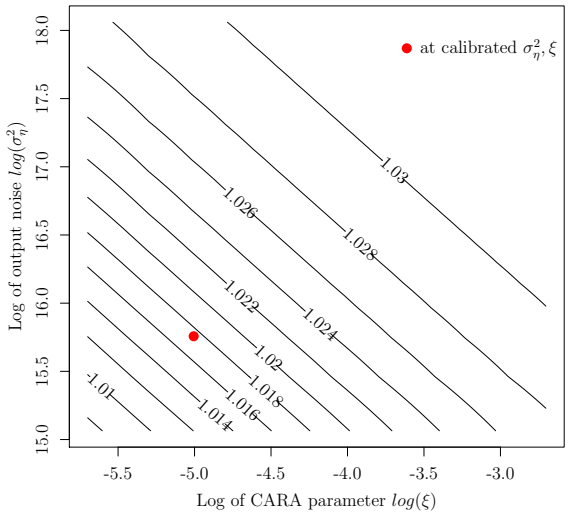
In contrast, using the above reckoning that empirical Bayes increases the variance of η by 3.2%, using empirical Bayes would produce an optimal slope of $\beta_1^{*EB} = 0.446$ and optimal action of $a^{*EB} = \$17,306.33$, i.e., a 1.045 sd average increase in student achievement. As expected, the higher measurement error variance on output from using empirical Bayes would lower the strength of incentives (i.e., slope) and resulting equilibrium action. Output would be 1.75% higher under fixed effects than it would be under empirical Bayes, suggesting an obvious choice of fixed effects for education policymakers. Naturally, we would expect the results from the hidden type model to be smaller than those from hidden action model here, as the hidden type model primarily affects output at the low end of the teacher quality distribution, while the hidden action model would affect output for all teachers.

Sensitivity Analysis Via Parameter Grid The mean class size in the Los Angeles data is 22.5, much smaller than the mean of 37.5 used in the above calibration. Smaller class sizes would increase the variance of the output measure. Moreover, Dohmen and Falk (2010) document that teachers are more risk-averse than other workers. Therefore, it would seem reasonable to examine how estimator-specific output would be affected by varying the parameters of the hidden action model. Mehta (2018) does this for a grid of points covering a wide range of alternative values of σ_η^2 and ξ, ranging from one half to ten times the calibrated value of each parameter.[41] Briefly, as teachers become more risk-averse (increasing ξ) or the output measure becomes noisier (increasing σ_η^2), both optimal incentive strength and output would decrease. For example, the increase in output ranges from over 1.5 sd in student achievement to around 0.5 sd when teachers are ten times more risk-averse than their calibrated value of $\xi = 6.7 \times 10^{-3}$; this latter figure is only about three times the estimated effect of the incentive scheme.

As interesting as these results may be in their own right, the goal here is to quantify the difference in output stemming from using one estimator versus another. Figure 6 presents a contour map of the ratio in optimal output from using fixed effects over that using empirical Bayes. Although, as just discussed, optimal incentive strength and output gains vary considerably with respect to σ_η^2 and ξ, the output gain associated with using fixed effects versus empirical Bayes ranges from just above 1% to around 3%. Intuitively, the higher noise in empirical Bayes matters more (relative to the cost γ) when teachers are more risk averse or when the baseline variance on the shock to output is higher. Of course, we cannot know the

[41] Table 2 in Babcock et al. (1993) shows that a higher-end estimate of ξ is about 0.35, well above the range considered in the parameter grid here. The output loss from using empirical Bayes would be larger for CARA parameters in that range. Note that, because γ was recovered using the teacher's optimal action choice and can be recovered by using the slope of incentives in the experiment and increase in output, it does not depend on (σ_η^2, ξ) and is therefore fixed.

Figure 6: Ratio of optimal output,
$\mathrm{E}\left[q^{*FE}\right] / \mathrm{E}\left[q^{*EB}\right]$



This figure plots contours of the ratio of estimator-specific optimized quality (fixed effects over empirical Bayes), for a range of the CARA parameter ξ (the log of which is the x-axis) and variance of output noise σ_η^2 (the log of which is the y-axis). The calibrated values of ξ and σ_η^2 are depicted by the red dot.

exact amount by which the output would be lower were the administrator to use empirical Bayes; knowing this would require the development and estimation of a richer structural model, a promising avenue for future research. However, the variable share of compensation, calculated in Mehta (2018), can provide further guidance. As with the slope and output, this share declines as the output noise variance and degree of risk aversion increase. Suppose it seemed reasonable that, in the optimal arrangement, the variable share of compensation for teachers would be at most around 2% of their income. Then the gain in output from switching from empirical Bayes to fixed effects would be about 2-3%, which is even larger than it was at the calibrated parameter values.

6 Discussion

While economic theory can help inform education policy, measurement issues are also important when considering how to actually use data. Due to their statistical properties, 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, personnel decisions, or wages. More generally, shrinkage estimators are used on a broad array of policy-relevant applications. It is not obvious this

should be the case. If an estimator is going to be used to make a decision, then studies of its bias and other statistical properties are certainly useful, but as an intermediate—not final—step in their evaluation.

In this paper, I show that the preferred estimator depends on information that is plausibly part of an administrator's context. The preferred estimator would be the same for wide ranges of underlying parameters for all the models considered and is determined by the qualitative relationship between class size and teacher quality. I find that class size is negative-quadratic with respect to teacher quality in the Los Angeles Unified School District. Suppose an administrator had been using empirical Bayes in an incentive scheme. Would it make sense to switch to fixed effects? The relevant comparison, from an economic perspective, is a cost-benefit one. It is important to note that the intervention considered in this paper is very easy to implement and virtually costless—to use a different, more transparent estimator of teacher quality—and that the preferred estimator would be the same across several models of the administrator's objective. Indeed, in all likelihood, the relative cost of using fixed effects would be zero, or even negative, given the increased transparency of fixed effects, which could translate to a lower nonpecuniary cost incurred by society. Then, by an economic criterion, these results suggest an obvious benefit from using fixed effects instead of empirical Bayes in the design of teacher incentive schemes if, as was suggested previously, class size is negative quadratic in teacher quality in the relevant context. Administrators hesitant to implement incentive schemes may take comfort in at least knowing schemes were better-designed for their purposes, reducing the change of public backlash. It is important to note that the results are not just driven by the systematic relationship between class size and teacher quality (i.e., class sizes that are negative quadratic in teacher quality). Absent this relationship, empirical Bayes would still not return a higher value to the administrator, as uniform shrinkage would be undone by the administrator's optimizing behavior

Motivated by the quantitative results showing the choice of estimator can create differences in policy-relevant outcomes, I have reviewed existing incentive schemes, which are summarized in Appendix A. Most of the schemes use cutoff rules to assign bonuses and more than half base bonuses, in part, on value-added models of student achievement. Almost 90% of these use empirical Bayes estimators to calculate teacher quality. Strikingly, about one-fifth of the schemes do not even specify how student achievement is mapped into teacher bonuses. A corollary of this paper's results is that, because the choice of estimator matters, teacher incentive programs should clearly specify exactly how student achievement enters them.

This paper characterizes which estimator would be preferred by an administrator in an extremely large school district that has recently received much policy interest. A study of

how best to estimate teacher quality for another context would require data from the relevant geography and, to prescribe the optimal policy, information about the administrator's preferences. However, the uniform nature of the preferred estimator across the variety of environments studied in this paper suggests that a policymaker in another district could choose the right estimator for their context with a certain degree of confidence. Important future work would study optimal design of incentive schemes using a more general production technology model relating economic output to teacher quality, such as one allowing for cumulative effects of inputs in a dynamic setting.

This paper's findings could also in principle be applied to other work studying how to structure incentives and personnel decisions based on noisy output measures. For example, findings from the cutoff model could potentially be applied to decisionmakers in other settings that discrete policies (e.g., Rubin (1980), who uses empirical Bayes to study law school admissions decisions). The insights from this paper could also apply to other deviations from the statistical framework considered in this paper. For example, a biased estimator may be preferred in the cutoff model if higher-quality teachers were known to be systematically assigned unobservably better students, as this would dilate quality measures and make it easier to identify teachers of interest. Although deriving the (fully) optimal estimator for the environments considered in this paper is a technically difficult problem, future research making progress in this area could also quantify the effects of using such an estimator.

Macartney (2016) documents that teachers are forward-looking, choosing effort levels in response to current and future incentives. Therefore, another important extension would extend this paper's static framework, where the most natural starting point would likely be the hidden action model, to examine how measurement and optimal remuneration policy would interact to affect the administrator's value in dynamic settings. This would entail characterizing the optimal contract in a dynamic setting, and then comparing the value according to each estimator under this contract. Doing so would require considerable technical advances, as the optimal contract designed by the administrator would have to characterize the optimal sequence of functions in a dynamic setting, which is an active area of research (see, e.g., Clausen (2013), Piskorski and Westerfield (2016)). Moreover, such an extension would have to address the comparability of quality measures between periods (Bond and Lang (2013), Penney (2017)), a concern obviated by this paper's static approach. Therefore, this ambitious undertaking is left for future research.

Appendix

A Teacher Incentive Schemes

Table A.1 documents existing teacher incentive schemes that are based, at least in part, on student achievement. Many of these schemes include estimates of value-added as a determinant of teacher bonuses, and most that do base bonuses on value-added also include other measures of teacher quality.

Table A.1: Incentive pay schemes

Name of scheme	Location	Active dates	Bonus schedule	Uses value-added ?	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 value-added 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. A "?" means the information for that field was not available.

References

Andrabi, T., J. Das, A. I. Khwaja and T. Zajonc, "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics," *American Economic Journal: Applied Economics*, 3(3):29–54, 2011.

Angrist, J. D., P. D. Hull, P. A. Pathak and C. R. Walters, "Leveraging Lotteries for School Value-added: Testing and Estimation," *Quarterly Journal of Economics*, 132(2):871–919, 2017.

Athey, S., "Beyond Prediction: Using Big Data for Policy Problems," *Science*, 355(6324):483–485, 2017.

Babcock, B. A., E. K. Choi and E. Feinerman, "Risk and Probability Premiums for CARA Utility Functions," *Journal of Agricultural and Resource Economics*, pp. 17–24, 1993.

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

Barlevy, G. and D. Neal, "Pay for Percentile," *American Economic Review*, 102(5):1805–31, 2012.

Barrett, N. and E. F. Toma, "Reward or Punishment? Class Size and Teacher Quality," *Economics of Education Review*, 35:41–52, 2013.

Behrman, J. R., M. M. Tincani, P. E. Todd and K. I. Wolpin, "Teacher Quality in Public and Private Schools Under a Voucher System: The Case of Chile," *Journal of Labor Economics*, 34(2):319–362, 2016.

Bolton, P. and M. Dewatripont, *Contract Theory*, MIT Press, 2005.

Bond, T. N. and K. Lang, "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.

Buddin, R., "Measuring Teacher and School Effectiveness at Improving Student Achievement in Los Angeles Elementary Schools," *RAND Corporation Working Paper*, 2011.

Chambers, C. P. and P. J. Healy, "Updating Toward the Signal," *Economic Theory*, 50(3):765–786, 2012.

Chetty, R., J. N. Friedman and J. E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104(9):2593–2632, 2014a.

—, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104(9):2633–2679, 2014b.

—, "Measuring the Impacts of Teachers: Reply," *American Economic Review*, 107(6):1685–1717, 2017.

Chetty, R. and N. Hendren, "The Impacts of Neighborhoods on Intergenerational Mobility: Childhood Exposure Effects and County-Level Estimates," *Harvard University and NBER*, pp. 1–144, 2016.

Clausen, A., "Moral Hazard with Counterfeit Signals," *SIRE Discussion Paper*, SIRE-DP-2013-13, 2013.

Clotfelter, C. T., H. F. Ladd and J. L. Vigdor, "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources*, 41(4):778–820, 2006.

Cohen, A. and L. Einav, "Estimating Risk Preferences From Deductible Choice," *American Economic Review*, pp. 745–788, 2007.

Copas, J. B., "Regression, Prediction and Shrinkage," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 311–354, 1983.

Ding, W. and S. F. Lehrer, "Estimating Treatment Effects From Contaminated Multiperiod Education Experiments: The Dynamic Impacts of Class Size Reductions," *Review of Economics and Statistics*, 92(1):31–42, 2010.

—, "Understanding the Role of Time-Varying Unobserved Ability Heterogeneity in Education Production," *Economics of Education Review*, 40:55–75, 2014.

Dohmen, T. and A. Falk, "You Get What You Pay For: Incentives and Selection in the Education System," *Economic Journal*, 120(546):F256–F271, 2010.

Ferrall, C. and B. Shearer, "Incentives and Transactions Costs Within the Firm: Estimating an Agency Model Using Payroll Records," *Review of Economic Studies*, 66(2):309–338, 1999.

Fryer Jr. R. G. and S. D. Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School," *Review of Economics and Statistics*, 86(2):447–464, 2004.

Glazerman, S., S. Loeb, D. Goldhaber, D. Staiger, S. Raudenbush and G. Whitehurst, "Evaluating Teachers: The Important Role of Value-Added," Tech. rep., Mathematica Policy Research, 2010.

Goldstein, D., "Randi Weingarten: Stop the Testing Obsession," *Dana Goldstein's Blog at The Nation*, 2012.

Greene, W. H., *Econometric Analysis*, Prentice Hall, Upper Saddle River, New Jersey 07458, 5th edn., 2003.

Guarino, C., M. Reckase and J. Wooldridge, "Can Value-Added Measures of Teacher Performance Be Trusted?" *Education Finance and Policy*, 10(1):117–156, 2014.

Guarino, C. M., M. Maxfield, M. D. Reckase, P. N. Thompson and J. M. Wooldridge, "An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures," *Journal of Educational and Behavioral Statistics*, 40(2):190–222, 2015.

Hansen, K. T., J. J. Heckman and K. J. Mullen, "The Effect of Schooling and Ability on Achievement Test Scores," *Journal of Econometrics*, 121(1):39–98, 2004.

Hanushek, E. A., "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, pp. 351–388, 1979.

—, "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24(3):1141–1177, 1986, ISSN 0022-0515.

—, "The Economic Value of Higher Teacher Quality," *Economics of Education Review*, 30(3):466–479, 2011.

Hölmstrom, B., "Moral Hazard and Observability," *Bell Journal of Economics*, pp. 74–91, 1979.

Hölmstrom, B. and P. Milgrom, "Aggregation and Linearity in the Provision of Intertemporal Incentives," *Econometrica*, pp. 303–328, 1987.

Imberman, S. A. and M. F. Lovenheim, "Does the Market Value Value-Added? Evidence from Housing Prices After a Public Release of School and Teacher Value-Added," *Journal of Urban Economics*, 91:104–121, 2016.

Jackson, C. K., "Teacher Quality at the High School Level: The Importance of Accounting for Tracks," *Journal of Labor Economics*, 32(4):pp. 645–684, 2014, ISSN 0734306X.

Jepsen, C. and S. Rivkin, "Class Size Reduction and Student Achievement: The Potential Tradeoff Between Teacher Quality and Class Size," *Journal of Human Resources*, 44(1):223–250, 2009.

Jepsen, C. and S. G. Rivkin, *Class Size Reduction, Teacher Quality, and Academic Achievement in California Public Elementary Schools*, Public Policy Institute of California, 2002.

Kane, T. J., J. E. Rockoff and D. O. Staiger, "What Does Certification Tell Us About Teacher Effectiveness? Evidence From New York City," *Economics of Education Review*, 27(6):615–631, 2008.

Karlin, S. and H. Rubin, "Distributions Possessing a Monotone Likelihood Ratio," *Journal of the American Statistical Association*, 51(276):637–643, 1956.

Kinsler, J., "Assessing Rothstein's Critique of Teacher Value-Added Models," *Quantitative Economics*, 3(2):333–362, 2012a.

—, "Beyond Levels and Growth: Estimating Teacher Value-Added and its Persistence," *Journal of Human Resources*, 47(3):722–753, 2012b.

—, "Teacher Complementarities in Test Score Production: Evidence from Primary School," *Journal of Labor Economics*, 34(1):29–61, 2016.

Koedel, C. and J. Betts, "Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation," *Education Finance and Policy*, 5(1):54–81, 2010.

Lazear, E. P., "Educational Production," *Quarterly Journal of Economics*, pp. 777–803, 2001.

Lippman, S. A. and J. McCall, "The Economics of Job Search: A Survey," *Economic Inquiry*, 14(2):155–189, 1976.

Macartney, H., "The Dynamic Effects of Educational Accountability," *Journal of Labor Economics*, 34(1):1–28, 2016.

Macartney, H., R. McMillan and U. Petronijevic, "A Unifying Framework for Education Policy Analysis," *NBER Working Paper*, 2016.

Makov, U. E., A. F. Smith and Y.-H. Liu, "Bayesian Methods in Actuarial Science," *The Statistician*, pp. 503–515, 1996.

Manski, C. F., "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72(4):1221–1246, 2004.

McCaffrey, D. F., J. Lockwood, D. M. Koretz and L. S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability. Monograph.*, ERIC, 2003.

McCaffrey, D. F., T. R. Sass, J. Lockwood and K. Mihaly, "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy*, 4(4):572–606, 2009.

Mehta, N., "The Potential Output Gains from Using Optimal Teacher Incentives: An Illustrative Calibration of a Hidden Action Model," *Economics of Education Review*, 66:67–72, 2018, doi:10.1016/j.econedurev.2018.06.011.

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

Muralidharan, K. and V. Sundararaman, "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, 119(1):39–77, 2011.

Murnane, R. J., *The Impact of School Resources on the Learning of Inner City Children.*, Cambridge, MA: Ballinger, 1975.

Oakes, J. M., "The (Mis) Estimation of Neighborhood Effects: Causal Inference for a Practicable Social Epidemiology," *Social Science & Medicine*, 58(10):1929–1952, 2004.

Papay, J. P. and M. A. Kraft, "Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-term Career Improvement," *Journal of Public Economics*, 130:105–119, 2015.

Penney, J., "Test Score Measurement and the Black-White Test Score Gap," *Review of Economics and Statistics*, 99(4):652–656, 2017.

Piskorski, T. and M. M. Westerfield, "Optimal Dynamic Contracts with Moral Hazard and Costly Monitoring," *Journal of Economic Theory*, 166:242–281, 2016.

Player, D., "Nonmonetary Compensation in the Public Teacher Labor Market," *Education Finance and Policy*, 5(1):82–103, 2010.

Raudenbush, S. and A. S. Bryk, "A Hierarchical Model for Studying School Effects," *Sociology of Education*, pp. 1–17, 1986.

Rivkin, S. G., E. A. Hanushek and J. F. Kain, "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," *American Economic Review*, 94(2):247–252, 2004.

Rossi, P. H., M. W. Lipsey and H. E. Freeman, *Evaluation: A Systematic Approach*, Sage Publications, 2003.

Rothstein, J., "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy*, 4(4):537–571, 2009.

—, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics*, 125(1):175–214, 2010.

—, "Teacher Quality Policy When Supply Matters," *American Economic Review*, 105(1):100–130, 2014.

—, "Measuring the Impacts of Teachers: Comment," *American Economic Review*, 107(6):1656–1684, 2017.

Rubin, D. B., "Using Empirical Bayes Techniques in the Law School Validity Studies," *Journal of the American Statistical Association*, 75(372):801–816, 1980.

Schochet, P. and H. Chiang, "What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models?" *Journal of Educational and Behavioral Statistics*, 2012.

Shiryaev, A. N., *Optimal Stopping Rules*, vol. 8, Springer Science & Business Media, 2007.

Staiger, D. and J. Rockoff, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives*, 24(3):97–117, 2010.

Stiglitz, J. E., "Symposium on Organizations and Economics," *Journal of Economic Perspectives*, pp. 15–24, 1991.

Stinebrickner, T. R., "A Dynamic Model of Teacher Labor Supply," *Journal of Labor Economics*, 19(1):196–230, 2001.

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 Journal of Educational Research*, 42:1–21, 2004.

The Sage Developers, *SageMath, the Sage Mathematics Software System (Version 8.1)*, 2017, `http://www.sagemath.org`.

Tincani, M. M., "Teacher Labor Markets, School Vouchers and Student Cognitive Achievement: Evidence from Chile," Ph.D. thesis, University of Pennsylvania, 2012.

Todd, P. and K. Wolpin, "The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps," *Journal of Human Capital*, 1(1):91–136, 2007.

Todd, P. E. and K. I. Wolpin, "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal*, 113(485):F3–F33, 2003.

—, "Estimating a Coordination Game in the Classroom," *Working Paper*, 2012.

Turque, B., "Rhee Dismisses 241 D.C. Teachers; Union Vows to Contest Firings," *The Washington Post*, 2010.

Wiswall, M., "The Dynamics of Teacher Quality," *Journal of Public Economics*, 100:61–78, 2013.