Targeting the Wrong Teachers? Linking Measurement with Theory to Evaluate Teacher Incentive Schemes

by

Nirav Mehta

Working Paper # 2017-1 **March 2017**

Centre for Human Capital and Productivity (CHCP)

Working Paper Series

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

Targeting the Wrong Teachers?

Linking Measurement with Theory to Evaluate Teacher Incentive Schemes

Nirav Mehta
University of Western Ontario *

March 28, 2017

Abstract

Measurement is crucial to the implementation of output-based incentive schemes. This paper uses models to study the performance of teacher quality estimators that enter teacher incentive schemes. I model an administrator tasked with (i) categorizing teachers with respect to a cutoff, (ii) retaining teachers in a hidden type environment, and (iii) compensating teachers in a hidden action environment. The preferred estimator would be the same in each model and depends on the relationship between teacher quality and class size. I use data from Los Angeles to show that simple fixed effects would almost always outperform more popular empirical Bayes.

Keywords: teacher incentive pay, teacher quality

*nirav.mehta@uwo.ca

1 Introduction

The vast majority of teacher remuneration is based on credentials and experience (Podgursky and Springer (2006, 2011)). However, the fact that only a small amount of variation in student achievement is explained by these characteristics (Hanushek (1986), Goldhaber and Brewer (1997), Rivkin et al. (2005)) and evidence that teacher quality is an important determinant of human capital (Hanushek (2011), Chetty et al. (2014a)) have spurred a debate about introducing teacher incentive pay schemes. Such schemes typically take as inputs estimates of teacher quality,[1] prompting a seemingly simple, but clearly germane question: How should teacher quality be measured in the context of teacher incentive schemes?

At a broad level, the how to best measure quality depends on whether the goal is to *predict* teacher quality or to *maximize an economic objective*. The unbiased sample mean of measured teacher quality is an obvious candidate. However, the posterior mean of measured teacher quality is desirable for prediction because it minimizes mean squared error. It accomplishes this by downweighting, or "shrinking", the sample mean toward the population mean, which introduces bias while reducing variance. Lacking any other information about the economic environment, minimizing mean squared error may seem like a reasonable criterion. Indeed, this may explain the ubiquity of shrinkage estimators in research and practice (e.g., Rockoff (2004) and Kane et al. (2008)).[2] However, which estimator would be preferred would likely depend on an educational administrator's context, which we may know something about. It is not clear that an estimator with a lower mean squared error would be preferred over one that is unbiased for all relevant environments. Incentive schemes are predicated on the economic theory of incentives. As such, it seems that estimators of teacher quality should be evaluated on the basis of economic—not statistical—theory, which takes into account the relevant context.

This paper combines economic theory and data to examine how to best estimate teacher quality in the context of a utility maximizing entity, such as a school district administrator. I consider the ubiquitous "empirical Bayes" estimator, a type of shrinkage estimator,[3] and the unbiased fixed effects estimator, i.e., the sample mean of teacher quality net of other observed inputs. The aforementioned bias-variance tradeoff emerges whenever comparing unshrunk estimates with shrunken ones. The framework developed in this paper could also be used to examine the performance of other estimators.

I consider the choice of estimator in three economic environments that are salient for ed-

[1] I refer to "quality" and "value-added" interchangeably.

[2] McCaffrey et al. (2003) write "Early [value-added model] applications ... primarily used fixed effects, while more recent applications ... have used random effects almost exclusively" (64). Note that though empirical Bayes can be viewed as the outcome of a random effects model, it also possible to shrink fixed effects using a Bayesian approach.

[3] Empirical Bayes estimators are often called BLUP (Best Linear Unbiased Predictor) because they minimize the mean squared error.

ucation policy. To most closely match the structure of the vast majority of existing schemes, I start by developing a cutoff-based model where the administrator chooses an optimal cutoff policy to classify teachers with respect to a desired threshold quality, minimizing the weighted sum of expected Type I and Type II errors. Each policy specifies a cutoff in the distribution of that estimator say, for receiving a bonus or not being dismissed. Intuitively, her expected utility is equal to the expected probability of correct classifications, making it natural to compare estimators based on their expected frequency of mistakes.

The estimators differ by how much they weigh sample data. In the context of estimating teacher quality, these weights increase in class size. Analysis of the cutoff model shows that the relationship between class size and teacher quality determines which estimator the administrator would prefer. The administrator obtains the same expected maximized objective (or "value") when class size is constant because empirical Bayes shrinks estimates for all teachers towards the population mean by the same proportion, preserving teacher rankings. However, if school principals shift students away from the lowest-quality teachers or assign the highest-quality teachers to teach small classes of gifted students then class size may depend on teacher quality. This would cause the performance of the estimators to diverge, even when the administrator uses estimator-specific optimal cutoff policies. To see why, consider a comparison between two teachers of different below-average qualities. If the higher-quality teacher is assigned more students, the larger number of signals about teacher quality increases the empirical Bayes' weight on their students' test score gains and decreases the weight on the population mean of teacher quality, relative to the lower-quality teacher assigned fewer students. In the extreme scenario where all but one student in a large school are assigned to the higher-quality teacher, both estimators for that teacher converge to the true value. However, the empirical Bayes estimate for the lower-quality teacher will likely be close to the population mean, implying that an administrator using empirical Bayes would likely determine that the lower-quality teacher is better. I show that the performance of the estimators differs most at the tails of the distribution of teacher quality under several plausible scenarios, which is important if we seek to identify either high- or low-quality teachers. For example, in 2010, the former Washington D.C. Schools Chancellor Michelle Rhee fired 241 teachers based on performance measures (Turque (2010)).

The cutoff model's close link to existing policy is clearly desirable. However, the cutoff model does not directly link measurement and output. Therefore, I also use the two main types of asymmetric information models to directly study how measurement of teacher quality may affect output. In each model, the administrator chooses an estimator-specific reward policy function to maximize her objective: output net the cost of the policy. The first is a hidden type, or adverse selection, model in which unobserved types determine teacher quality. Teacher quality is a determinant of student human capital and, hence, output, but cannot be exactly recovered due to measurement error. I show that a reservation-value-, i.e., cutoff-, based policy

would emerge in this environment. The optimality of a reservation-value policy suggests a link between the cutoff and hidden type models. Indeed, I show that the administrator's preferred estimator in the hidden type model also depends on class size scenario and, moreover, would be the same as her preferred estimator in the cutoff model. When class sizes are constant she would be indifferent, when they are negative quadratic in teacher quality she would prefer fixed effects, and the opposite when they are positive quadratic. Intuitively, in both the cutoff and hidden type models the administrator's value is higher the easier it is to identify teachers above a particular threshold, and lower the more likely it is that high-quality teachers are estimated to be below (or low-quality teachers are found to be above) that threshold.

The second asymmetric information model is a hidden action, or moral hazard, model, based on Hölmstrom and Milgrom (1987). In this model, teachers take an unobserved action which determines their quality. As in the hidden type model, teacher quality affects output but cannot be exactly recovered because output is measured with noise. Hölmstrom and Milgrom (1987) show that the optimal incentive scheme in this environment is linear in the output signal, which depends on estimated teacher quality. As with the other models, I show that the administrator would be indifferent between the estimators when class size is constant. I also show that the administrator can change the scale of the estimator and achieve the same utility. This means the key link between measurement and choice of estimator is that a negative-quadratic (positive-quadratic) relationship between class size and teacher quality can be modeled as a larger (smaller) noise component to an administrator paying teachers using the optimal contract based on the empirical Bayes estimator. A larger noise component would reduce the strength of optimal incentives, i.e., piece rate, if agents are risk averse, as teachers likely are. This model provides an intuitive economics-based response to the sentiment that there is "too much" noise in teacher quality measures[4]: This "big" variance would, in equilibrium, result in a flatter optimal wage schedule. "Changing the data" by, e.g., shrinking fixed effects estimates would only lead to an improvement if class size were positive quadratic in teacher quality. Therefore, as in the cutoff and hidden type models, the administrator's preferred estimator depends on the class size scenario and, as in the hidden type model, lines up with her preferred estimator in the cutoff model.

I also show that the administrator's preferred estimator would be the same for a much more general objective, which is increasing in the product of teacher quality (or monotonic transformation thereof) and the reward assigned to the teacher, where the reward has the natural property of being nondecreasing in estimated quality.

A distinguishing feature of this paper, relative to the literature, is that for each model I con-

[4]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)).

sider the administrator chooses an optimal reward policy function for each estimator. Moreover, for each model, the administrator's optimized expected utility, or value, according to either estimator is characterized for a wide range of underlying parameters. This approach answers a different type of question than one quantifying the effects of potentially suboptimal policies using estimated models (e.g., Stinebrickner (2001), Tincani (2012), Todd and Wolpin (2012), Behrman et al. (2016)) or even those with calibrated parameters (e.g., Rothstein (2014)). The advantage of the approach taken here is that it is possible to compute the preferred estimator—i.e., the one returning a higher value when coupled with the estimator-specific optimal reward policy—*without knowing the specific parameterization of the relevant model*. All that is required is the relationship between class size and teacher quality. This is feasible because in each environment I consider, the administrator's problem can be split into two parts: choose a teacher quality estimator and then choose an estimator-specific optimal policy.

Given the theoretical results, the natural next step is to establish which estimator would likely be preferred in the real world. Moreover, can we—at least roughly—quantify the extent to which the choice of estimator matters? As the models show, the former question boils down to the relationship between class size and teacher quality, which, to my knowledge, has not been well-established. Therefore, I first make an empirical contribution by documenting this relationship for the Los Angeles Unified School District—the second-largest school district in the United States and a district with a large degree of diversity and variation in both student achievement and class size—using value-added estimates provided by the Los Angeles Times (Buddin (2011)). I find that class size increases in teacher quality at the low end of the quality distribution and decreases in teacher quality at the high end; Section 2 shows there is reason to believe similar relationships between class size and teacher quality may also be present elsewhere. This is the class size scenario under which fixed effects would be preferred over empirical Bayes in all the models. However, because the value to the administrator and gain in output depend on model primitives, without further information this approach cannot quantify the magnitude of the change according to using one estimator over another.

To address this limitation, I then use the models to compare the prospective performance of the estimators. First, to approximate the objective of an administrator considering implementing a district-wide cutoff-based incentive scheme, I use values calibrated from Schochet and Chiang (2012) and the now-documented relationship between teacher quality and class size in Los Angeles to solve for optimal cutoff policies for a wide range of prospective desired cutoffs for each estimator. Because the cutoff model has few parameters, I can solve for the preferred estimator for every desired cutoff, obviating calibrating additional parameters. Fixed effects would perform better than empirical Bayes for almost every desired cutoff because the empirical relationship I document between class size and teacher quality would cause empirical Bayes to correctly place fewer teachers of extreme quality in the tails, making it harder to separate

them from other teachers. For example, I find that the administrator would make 10% more classification errors by switching from fixed effects to empirical Bayes to categorize teachers as being in the bottom percentile in Reading value-added in the Los Angeles Unified School District. The recent outcry about a case where value-added was incorrectly calculated for 40 teachers in Washington DC, which resulted in at least one firing (Strauss (2013)), suggests that the public is concerned about misclassifying public school teachers.

Next, I calibrate the additional parameters required to obtain a rough assessment of how the choice of estimator would affect output in the asymmetric information models. For the hidden type model, I calibrate a value for the cost of replacing a teacher and compute that using fixed effects instead of empirical Bayes would increase output by around 0.11-0.22% in a period. For the hidden action model, I calibrate model parameters using results from Muralidharan and Sundararaman (2011), an experimental study of teacher incentive pay implemented in Andhra Pradesh. I find that using fixed effects instead of empirical Bayes would increase output by 1.65% per period. This finding requires making an ancillary, yet interesting, contribution: using the lens of the hidden action model to interpret the results of this teacher incentive pay experiment and solve for the optimal strength of incentives in a hidden action environment. Under fixed effects, the optimal slope of incentives in measured output would be more than six times higher than it was in the experiment, under the calibrated parameters.

This paper provides evidence that, despite its desirable statistical properties, the most popular estimator of teacher quality would be outperformed by simpler one. More generally, by combining knowledge about an administrator's context with an economic model, we can obtain much-needed guidance for how to best estimate quality for use in teacher incentive schemes. With appropriate information about administrator's preferences and the relationship between class size and teacher quality, the approach taken here could be used to evaluate the preferred estimator in other environments. The first steps taken in this paper provide a framework that could be used to compare the performance of other ways of measuring teacher quality, or other important economic variables, as well.

The rest of this paper is organized as follows. Section 2 provides background and discusses related literature. Section 3 develops and analyzes the cutoff model. Section 4 considers the asymmetric information models: Section 4.1 develops the hidden type model and Section 4.2 illustrates results for a hidden action model. Section 5 presents the quantitative results. Section 6 concludes. The Appendix documents a number of teacher incentive pay schemes and also contains proofs and further details about the quantitative results.

2 Background

Correctly providing incentives in an environment where teachers may vary in both inherent effectiveness and unobserved effort is a difficult contracting problem, which has caused existing schemes to adopt several simplifications. First, student test score gains are assumed to separately depend on teacher quality and other inputs, resulting in a value-added model (Hanushek (1979)). Second, schemes often take the form of cutoff rules that reward (punish) teachers with estimated value-added above (below) some cutoff. For example, Glazerman et al. (2011) document that about half of the performance-based schemes drawing on the Teacher Improvement Fund are based on cutoff rules and I find that even a higher share are cutoff-based in my analysis of existing teacher incentive schemes in Appendix A. Finally, as mentioned above, decisions are typically based on either fixed effects or empirical Bayes estimates.

Value-Added Models Value-added models are the workhorse of existing teacher incentive schemes and education research. Due to their pervasiveness, I examine how the most commonly used estimators of value-added perform when the underlying technology is consistent with a value-added model. Therefore, this paper has a different focus than research studying how effectively value-added models measure teacher quality (see, e.g., Baker and Barton (2010), Guarino et al. (2014), Glazerman et al. (2010), and McCaffrey et al. (2003)). Value-added models are a restricted form of a more general production technology for cognitive achievement (Todd and Wolpin (2003)), and many authors have tested these restrictions to determine whether they are good measures of teacher quality, with mixed results. First, some authors have compared estimates of teacher value-added with and without random assignment of students to teachers (Kane and Staiger (2008), Kane et al. (2013)) or with subjective ratings of teacher effectiveness (Jacob and Lefgren (2008)), surmising that value-added models do a reasonably good job of measuring teacher quality.[5] Second, there is also concern that value-added models do not condition on sufficiently rich information about other inputs (Rothstein (2009, 2010), Andrabi et al. (2011), Jackson (2014)), though the evidence is mixed here as well (Kinsler (2012a), Kinsler (2012b), Chetty et al. (2014a), Kinsler (2016)). Bond and Lang (2013) 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.[6] Related to this, Cawley et al. (1999) find a nonlinear (and non-log-linear) relationship between test scores and wages, which suggests that, consistent with the cutoff model, ranking teachers may be useful.

Although many studies examine the statistical validity of value-added models, none compare how estimators of value-added perform from the perspective of a utility-maximizing adminis-

[5] Of course, additional data could improve estimates of teacher quality. Teixeira-Pinto and Normand (2009) develop a method that combines data from binary and continuous variables that predict common outcomes.

[6] In this paper, comparisons are made within one academic year.

trator. Schochet and Chiang (2012) calculate error rates for fixed effects and empirical Bayes estimators of teacher quality, assuming the same cutoff policy for both estimators. 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.

Endogenous Class Size The idea that class size can reflect information about teacher quality has theoretical precedent and empirical support. Jacob and Lefgren (2008) and Lang (2010) argue that principals know who the good and bad teachers are. Lazear (2001) develops a theoretical model of class size and teacher quality. Barrett and Toma (2013) assume that higher quality teachers have a smaller reduction in efficacy for a given increase in class size. Consider then, a principal wanting to have students pass a low proficiency threshold and increase total output at her school. The former could cause class size to increase in teacher quality at the low end of her quality distribution. However, due to the lack of flexible wages in the public education sector, she may also reduce class size at the high end of the quality distribution to retain high-quality teachers.

Empirically, many papers have found a relationship between teacher value added and non-monetary aspects of remuneration. For example, Player (2010) documents that higher quality teachers have fewer black students, students with learning disabilities, and males—all characteristics related to how difficult it is to teach such students. Clotfelter et al. (2006) show that highly qualified teachers tend to be matched with more advantaged students. Jepsen and Rivkin (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.

Design of Teacher Incentive Schemes This paper also relates to the literature viewing teacher payment as a contracting problem, where the administrator chooses the contract that induces the most effort given that she observes only a noisy measure of output (e.g., student scores on standardized tests). Barlevy and Neal (2012) combine the earlier literature on tournaments (Lazear and Rosen (1981), Green and Stokey (1983)) and the multi-task problem of Hölmstrom and Milgrom (1991) to specifically study how best to make comparisons between teachers serving comparable groups of students.

To most closely match existing incentive schemes, the first part of this paper assumes the administrator follows a cutoff rule. However, I also show that such a rule could be used to implement a pay-for-percentile-type scheme (Barlevy and Neal (2012)) and, more importantly, would

naturally emerge as the optimal policy in a hidden type environment; the latter observation was also made by Staiger and Rockoff (2010).

3 Cutoff-Based Model

Overview A large body of empirical work evaluates value-added models and compares the statistical properties of fixed effects and empirical Bayes estimators. However, determining which would be the *preferred* estimator for making decisions about rewarding or punishing teachers requires an economic model that posits an objective function for a decision maker. In this section I develop a cutoff model, which formalizes the objective of a school-district administrator; characterizes her optimal cutoff policy; and shows the relationship among (i) how class size varies with teacher quality, (ii) her choice of estimator, and (iii) her expected maximized utility, i.e., value. To most closely match existing policies, she 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.

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, public education policy debates.[7] 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 applicable to the vast majority of existing teacher incentive pay schemes; as noted by Stiglitz (1991) and Ferrall and Shearer (1999), real-world incentive schemes are typically quite simple in structure–even when, in theory, they should depend on all observed signals (Hölmstrom (1979)). Fourth, related literature also considers cutoff-based policies, e.g., Staiger and Rockoff (2010), Hanushek (2011), Tincani (2012), Chetty et al. (2014b), and Rothstein (2014). Finally, the cutoff-based model is extremely flexible and does not require us to take a stand on what underlies variation in measured output, which could be heterogeneity in fixed teacher productivity types or unobserved actions. As such, it can capture relevant economic environments, e.g., as is shown below, one in which teachers are entered into a tournament (Barlevy and Neal (2012)).

[7] Section 4.1.1 explores similarities between the cutoff-based objective and optimal policy in a hidden type environment.

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 a true quality below κ (not making a Type II error). The administrator's utility from using the estimator $\hat{\theta}$ and cutoff policy c on a teacher of true quality θ is:

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

where α and $(1-\alpha)$ are her weights on not making Type I and II errors, respectively.[8] 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 has symmetric preferences, or values Type I and II errors equally.

Teacher quality is distributed according to $\theta_i \sim F = \mathrm{N}(0, \sigma_\theta^2)$, where F is known.[9] As discussed in Section 2, 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 $n(\theta)$.[10] 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. Note that what matters is the end relationship $n(\theta)$; 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 only adopt this spare technology to simplify model exposition; the quantitative results use value-added estimates that control for many characteristics. The fixed-effects (FE) estimator of θ_i is the sample mean, i.e., $\hat{\theta}_i^{FE} = \sum_j \frac{y_{ji}}{n_i} = \theta_i + \bar{\epsilon}_i$, and, 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

[8] I also analyze a version of the model where the administrator's objective is increasing in the distance between teacher quality and the cutoff. The administrator's preferred estimator would not change. Quantitatively, this change would inflate the performance difference between the estimators. Results are available upon request.

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

[10] If the number of students assigned to a teacher was a strictly monotonic function of teacher quality, teacher rankings could be perfectly recovered by comparing class sizes. Therefore, I assume in this section that the administrator cannot directly condition on class size; Appendix B.1 provides results for the case where the administrator may directly incorporate class sizes in her policy. There are two reasons to avoid this direct conditioning. Including class size would provide school principals with a direct incentive to manipulate class size, outside of any effects of class size on total output. Additionally, doing so would complicate the scheme, potentially reducing its attractiveness to policymakers

distribution is also normal, giving $\hat{\theta}_i^{EB} = \lambda_i \hat{\theta}_i^{FE} + (1-\lambda_i)\underbrace{\mathrm{E}[\theta]}_{0} = \lambda_i \hat{\theta}_i^{FE} = \lambda_i(\theta_i + \bar{\epsilon}_i)$, where $\lambda_i = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\epsilon^2/n_i}$ is the ratio of the true variation in teacher quality (the signal) relative to the estimated variation using the fixed effects estimator (the signal plus noise).[11] 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.[12] 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.

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:

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

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

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

For either estimator, an increase in the prospective cutoff policy c decreases the probability of correctly identifying a teacher with true quality above κ and increases the probability of correctly identifying a teacher with true quality below κ. The optimal cutoff 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,

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

[12] A common variant of the EB estimator estimates the overall mean of θ. If the overall mean of θ is not parametrized according to another distribution, the empirical Bayes estimator may not be deemed fully Bayesian.

c^{*EB} solves

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

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

Although the presentation of the cutoff model is for one classification problem, i.e., (α, κ), nothing precludes the administrator from performing multiple classification problems simultaneously, each with its own parameterization. For example, the model could also be used to implement a tournament-based scheme (e.g., "pay-for-percentile-type", as studied by Barlevy and Neal (2012)), by allowing for many discrete bonuses, one for each desired κ, which would correspond to percentiles of the distribution of teacher quality in the ensuing equilibrium. The bonus for each κ would be an increment above that for the κ immediately below.

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, in general, the administrator's value of the two estimators depends on the relationship between class size and teacher quality. The administrator's value also depends on her Type I and II error weights, α and $1-\alpha$, respectively. For simplicity, α has been set to 1/2; Appendix B.3 shows this does not drive the findings.

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 $\lambda(n(\theta)) = \lambda \in (0,1), \forall\theta$. Let c^{*FE} satisfy the administrator's first-order condition (3) when $\lambda = 1$. Because λ is constant, then $c^{*EB} = c^{*FE}\lambda$ also solves (3), and returns the same value (i.e., $v_{CP}^{FE}(\kappa) = v_{CP}^{EB}(\kappa)$). □

Figure 1 illustrates Proposition 1 by plotting the expected utility of the objective under the fixed effects estimator (solid red line) and the empirical Bayes estimator (dotted blue line)

as a function of the cutoff policy for each estimator (x-axis), assuming the same class size for all teachers. Each curve traces out the administrator's expected utility as a function of cutoff policies, given an exogenous desired cutoff quality κ. The left panel corresponds to a desired cutoff of the first percentile teacher, i.e., $\kappa = F^{-1}(0.01) < 0$, the middle panel corresponds to a desired cutoff of median teacher quality, i.e., $\kappa = F^{-1}(0.50) = 0$, and the right panel corresponds to a desired cutoff of the 95th percentile teacher, i.e., $\kappa = F^{-1}(0.95) > 0$. Extremely low or high cutoff policies cause both estimators to misclassify either all low- or high-performing teachers, respectively, which is why the administrator's expected utility is 1/2 at either extreme (recall $\alpha = 1/2$ in this example). 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. Because the curves for both estimators obtain the same maximum height in each panel, we can see that these are equal when class size does not vary by teacher quality (i.e., $\lambda(n(\theta))$ is constant). The utility-maximizing cutoff policy adjusts to take into account the larger variance of administrator utility under the fixed effects estimator. This is because if c^{*FE} $(= \frac{c^{*FE}}{1})$ solves (3), $|c^{*EB}|$ must be smaller than $|c^{*FE}|$ to satisfy equation (3), as $\lambda < 1$ for the empirical Bayes estimator. In the case where class size is constant, the optimal cutoff policies for both estimators are at the same quantiles of the estimator distributions; that is, the same share of teachers are rewarded under both estimators. For example, when the desired cutoff is the first percentile of teacher quality the cutoff policy that maximizes expected administrator utility when using fixed effects is below the optimal cutoff policy when using empirical Bayes (Figure 1a). In contrast, when the administrator desires to separate the top 5% (95th percentile) from the rest of teachers, the optimal cutoff policy under the fixed effects estimator is higher than that under the empirical Bayes estimator, again due to the larger variance of the fixed effects estimator for teacher quality (Figure 1c). However, this higher variance does not affect the administrator's value, or maximized expected utility (maximum height of each curve) because the administrator is risk neutral.[13]

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

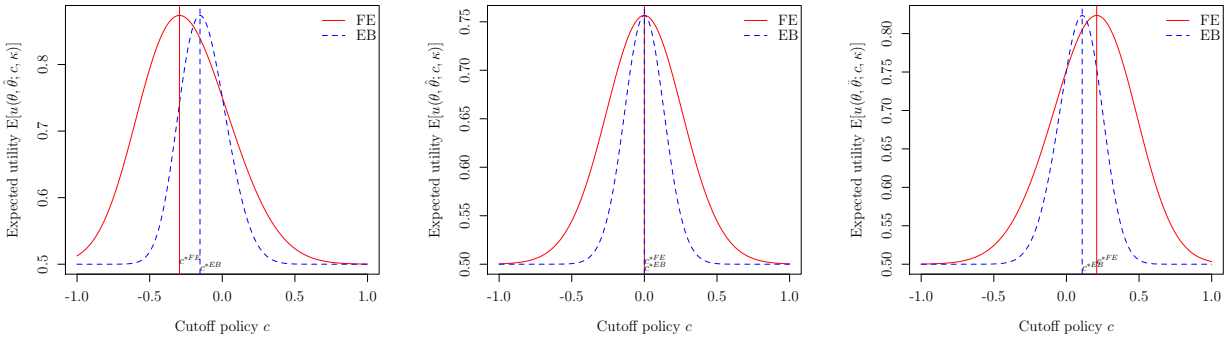
[13] The theoretical results, including Proposition 1, apply to deterministic class size functions; i.e., $n(\theta)$ is degenerate for each θ. If $\Phi(\cdot)$ were linear then the results would also apply for the case of i.i.d. class sizes. I have verified that the results, including estimator rankings, do not appreciably change when the administrator also integrates over i.i.d. class sizes; e.g., the administrator's objective under the fixed effects estimator is

$$\alpha \int_{\kappa}^{\infty} \left(\int_{\underline{n}}^{\overline{n}} \left(1 - \Phi\left(\frac{c^{FE} - \theta}{\sigma_\epsilon / n} \right) \right) dG_n(n) \right) dF(\theta | \theta \geq \kappa) + (1-\alpha) \int_{-\infty}^{\kappa} \left(\int_{\underline{n}}^{\overline{n}} \Phi\left(\frac{c^{FE} - \theta}{\sigma_\epsilon / n} \right) dG_n(n) \right) dF(\theta | \theta < \kappa)$$
$$= \alpha \int_{\kappa}^{\infty} 1 - \mathrm{E}_n \left[\Phi\left(\frac{c^{FE} - \theta}{\sigma_\epsilon / n} \right) \right] dF(\theta | \theta \geq \kappa) + (1-\alpha) \int_{-\infty}^{\kappa} \mathrm{E}_n \left[\Phi\left(\frac{c^{FE} - \theta}{\sigma_\epsilon / n} \right) \right] dF(\theta | \theta < \kappa),$$

where $G_n(\cdot)$ is a truncated normal random distribution chosen to fit the empirical distribution of class sizes. Due to the simpler exposition with degenerate class sizes this assumption is maintained. Results are available upon request.

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

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



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

Proof. Because λ is increasing in n, to simplify exposition I parametrize the empirical Bayes weights λ directly as a function of θ and then see how changes in this function would affect the administrator's utility from using the empirical Bayes estimator. In particular, I assume 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. I set $\sigma_{\bar{\epsilon}} = 1$ for all teachers for the proof of the current proposition, which does not affect the result; $\sigma_{\bar{\epsilon}}$ varies between teachers in the quantitative results. Parametrize the empirical Bayes weight according to

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

Suppose $\kappa < 0$ and that $c^{*EB} < 0$. The result holds if $\kappa > 0$ and $c^{*EB} > 0$, using analogous reasoning. Dividing through by $\alpha = 1/2$, the administrator's value is

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

Differentiate with respect to β_-:

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

where $\frac{\partial c^{*EB}}{\partial \beta_-} = 0$ due to the Envelope Theorem. The first term is negative because $-c^{*EB}\theta < 0$

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

Figure 2 illustrates Proposition 2 by plotting the administrator's objective under both estimators (equations (1) and (2)) against candidate cutoff policies (x-axis), but now under the assumption that class size is an increasing function of teacher quality, implying that $\beta_-, \beta_+ > 0$, meaning that lower-quality teachers are weighted closer to the population mean than higher-quality teachers. If the administrator desires to separate the lowest quality teachers from the rest (Figure 2a), the re-weighting inherent in the 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 2c)—the peak of the empirical Bayes curve is now higher than that under the fixed effects estimator. 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 2b), 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 3 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 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 $v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa)$

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

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

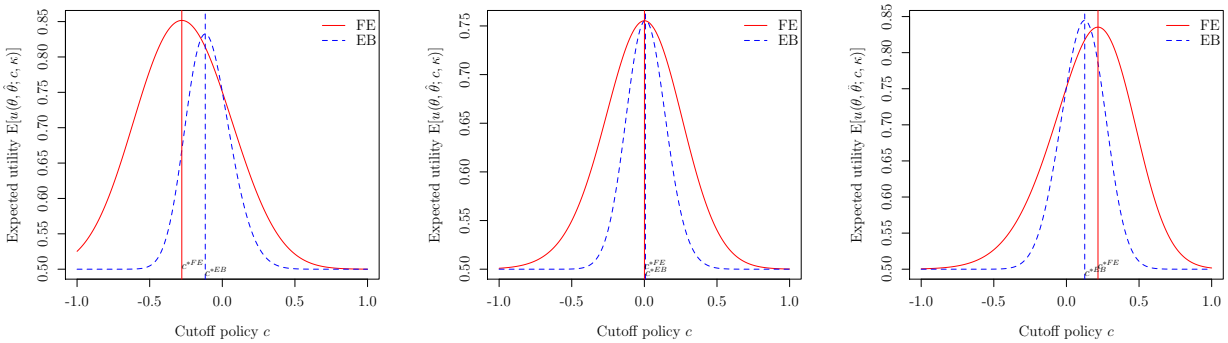
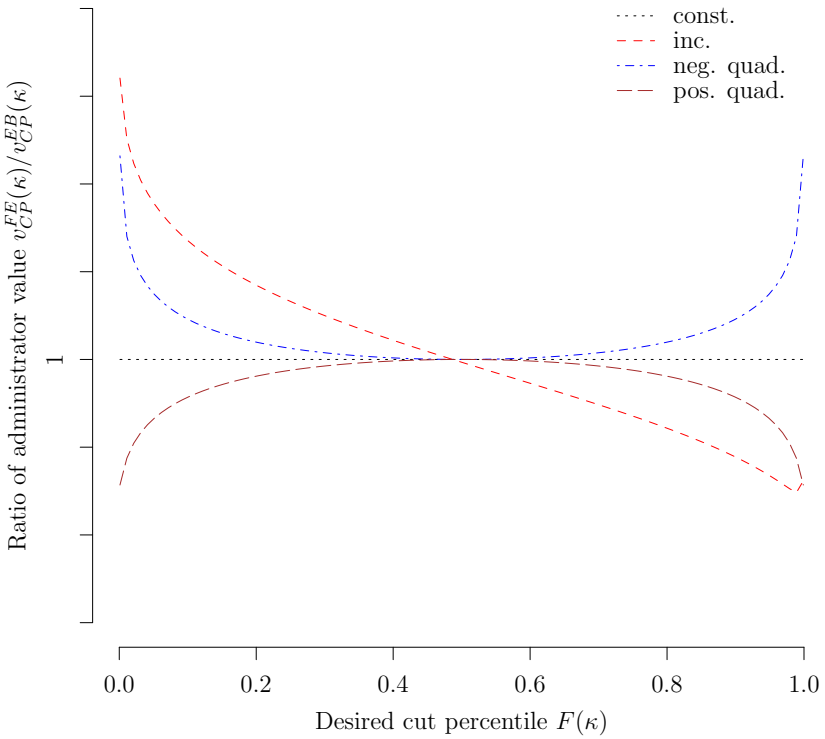


Figure 3: Difference between administrator's objective under fixed effects and empirical Bayes, by class size scenario and desired cut point



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 line), 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 2a), while the empirical Bayes estimator performs better when the administrator wishes to isolate high-quality teachers (Figure 2c). When class size has a negative-quadratic relationship with teacher quality (dot-dashed blue line), similar to the case in Proposition 2 where $\beta_- > 0$ and $\beta_+ < 0$, it is increasing when teacher quality is low and decreasing when teacher quality is high; in the example considered in Figure 3, the fixed effects estimator outperforms the empirical Bayes estimator at both the lowest and highest desired cutoffs. Finally, when class size is a positive-quadratic function of teacher quality (long-dashed brown line), the opposite is true. Figure 3 also demonstrates that the difference between the performance of 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 7 in Appendix B.4 for a proof that the administrator would be indifferent if her problem is *symmetric*).

It is important to note that, though the exposition here compares the performance of (unshrunken) fixed effects and (shrunken) empirical Bayes estimators, other Bayesian estimators could be accommodated by simply changing the prior variance σ_θ, which would affect the amount by which the sample mean is shrunk towards the mean.

As noted previously, the cutoff model could be applied to a tournament-based scheme, e.g., "pay-for-percentile" (Barlevy and Neal (2012)). Therefore, Proposition 2 shows that the administrator's objective in such an environment would be lower when using empirical Bayes when the relationship between class size and teacher quality is negative quadratic, the same if class sizes were constant, and higher when class size is positive quadratic in teacher quality. Intuitively, tournament-based schemes rely on ranking teachers, which is harder to do when lower- and higher-quality teachers are disproportionately shrunk towards the population mean (i.e., class size is negative quadratic in teacher quality).

4 Asymmetric Information Models

4.1 Hidden Type Model

This section shows how results from the cutoff-based model studied in Section 3 can be thought of in terms of a hidden type, or adverse selection, environment.[14] 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 the cutoff model the administrator was assumed to follow a cutoff policy. In contrast, this section shows that such a 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 conditions on quality signals, but not directly on other data, e.g., class sizes.[15] 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 $\underline{w}$. As in Staiger and Rockoff (2010), there is an arbitrarily large number of teachers for each slot. This is not very restrictive because a change in the distribution of teacher quality could be modeled by suitably adjusting the distribution of θ.

Teacher quality is not 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 $G_{\hat{q}}$ satisfies the Monotone Likelihood Ratio Property (MLRP), which is consistent with many distributions of measurement error on output—in particular, normally distributed errors (Karlin and Rubin (1956)), which are ubiquitous in value-added models. 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 $|H_{it}|$.

[14] This environment is partially based on one developed in Staiger and Rockoff (2010). See page 2 of their Online Appendix.

[15] Appendix B.1 examines a problem where the administrator can directly use class size in her policy.

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

$$u_{HTG} = \sum_t \delta^{t-1} \mathrm{E}_t \left[\left(\sum_j q_{i(j,t),t} - w_{i(j,t)} - 1\{|H_{i(j,t),t}| = 0\}\chi \right) \right], \tag{5}$$

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 $\underline{w} = 0$.[16] 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 $G_{\hat{q}}$ satisfies the MLRP, a continuation region above.*

Proof. First, note that the additive separability of (5) implies we can split it into J separate problems. Lippman and McCall (1976) proves that the optimal policy for each problem has a reservation value property (see also Rothschild (1974)). Examination of (5) shows that the administrator's objective is increasing in output q_{it}, and therefore also increasing in expected output. If the MLRP holds, 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 signal $\hat{q}_{it}$. Then, there will then be a 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 retention decision will have a reservation value property. Also note that variation in n_j does not affect the optimality of a reservation value policy, provided $G_{\hat{q}}$ satisfies the MLRP. □

[16] 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.

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. However, the administrator's objective (5) 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 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 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 signal $\hat{q}$ is

$$\underbrace{1\{\hat{q} < \underline{q}\}\left(\mathrm{E}\left[q|\text{new hire}\right] - \chi\right)}_{\text{dismiss teacher; fill slot immediately}} + \underbrace{1\{\hat{q} \geq \underline{q}\}\,\mathrm{E}\left[q|\hat{q} \geq \underline{q}\right]}_{\text{retain teacher}} = 1\{\hat{q} < \underline{q}\}(\underbrace{\mathrm{E}\left[\theta|\text{new hire}\right]}_{=\mu=0} - \chi) + 1\{\hat{q} \geq \underline{q}\}\,\mathrm{E}\left[\theta|\hat{q} \geq \underline{q}\right]. \tag{6}$$

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

$$v_{HT0}^{\hat{q}}(\chi) = \max_{\underline{q}} \Phi\left(\frac{\underline{q}}{\sigma_{\hat{q}}}\right)(-\chi) + \left(1 - \Phi\left(\frac{\underline{q}}{\sigma_{\hat{q}}}\right)\right)\mathrm{E}\left[\theta|\hat{q} \geq \underline{q}\right]. \tag{7}$$

By setting $\hat{q} = \hat{\theta}^{FE}$, the sample mean of each teacher's observed signals during the first

period, we can then use (7) to write the administrator's value from using the fixed effects estimator:

$$v_{HT0}^{FE}(\chi) = \max_{\underline{q}^{FE}} \Phi\left(\frac{\underline{q}^{FE}}{\sigma_{\hat{\theta}^{FE}}}\right)(-\chi) + \left(1 - \Phi\left(\frac{\underline{q}^{FE}}{\sigma_{\hat{\theta}^{FE}}}\right)\right)\frac{\sigma_\theta^2}{\sigma_{\hat{\theta}^{FE}}}\frac{\phi(-\underline{q}^{FE}/\sigma_{\hat{\theta}^{FE}})}{\Phi(-\underline{q}^{FE}/\sigma_{\hat{\theta}^{FE}})}, \tag{8}$$

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 could solve for the reservation signal $\underline{q}^{*FE}$ by differentiating (8) with respect to $\underline{q}^{FE}$ and setting the resulting first-order condition to zero. However, we can also 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 $\mathrm{E}[\theta] - \chi = -\chi$ 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 $-\frac{\chi}{\rho} \equiv \underline{q}^{*FE} > \hat{\theta}_{i(j,1)}^{FE}$. This expression has a clear intuition. First, suppose that $\chi = 0$. Then 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

The cutoff-based model in Section 3 has the advantage of being simple and embedding the administrator's objective in an intuitive, policy-relevant measure: the weighted sum of classification errors. 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.

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 (7) 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 (8) 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 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. Specifically, we can map Model HT-1 to a version of Model HT-0. This is formalized in Proposition 5.

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

Proof. See 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 (see Appendix C.2 for details). As was the case for the cutoff model, when class sizes are variable the preferred estimator depends on other primitives. Proposition 3 implies that the optimal policy in HT-2 would still have the reservation value property, so as long as the MLRP is maintained. This, combined with the fact that the administrator would prefer fixed effects to empirical Bayes for almost every parameterization of the cutoff model when $n(\theta)$ is

negative quadratic and would prefer the opposite when $n(\theta)$ is positive quadratic (Proposition 2), suggests the administrator would also prefer fixed effects in model HT-2 under the negative-quadratic scenario and would prefer empirical Bayes under the positive-quadratic scenario; i.e., the same ranking over estimators by class size scenario would also obtain for HT-2.

I confirmed this intuition by computing the preferred estimator for a range of parameterizations of Model HT-2, i.e., replacement costs χ. For brevity, the explicit model and results are presented in Appendix C.2. The results are strikingly similar between the cutoff and hidden type models: (i) which estimator the administrator would prefer depends on $n(\theta)$ (same as in the cutoff model), (ii) the preferred estimator does not depend on the specific parameterization of HT-2, other than the shape of $n(\theta)$ (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. That is, I find that the preferred estimator in the cutoff model, which depends on the class size scenario $n(\theta)$, would also be preferred in model HT-2, regardless of the values of other model parameters. Naturally, we might 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.

There is an intuition for why the administrator would prefer the same estimator in the cutoff and hidden type models. The administrator will have a higher value when there are fewer Type I errors, i.e., teachers with high true quality appearing below the reservation signal, because the cost of replacing them will be lower. At the same time, fixing the share of teachers not retained, the administrator will have a higher objective coming from fewer Type II errors, i.e., teachers with low true quality appearing above the reservation signal, because output will be higher. A negative-quadratic relationship between teacher quality and class size would, therefore, lead the administrator to prefer to use fixed effects over empirical Bayes in both the cutoff model and Model HT-2. Appendix E shows that an administrator with a much more general objective would also prefer fixed effects over empirical Bayes when $n(\theta)$ was negative quadratic and empirical Bayes when $n(\theta)$ was positive quadratic.

4.2 Hidden Action Model

As discussed previously, many teacher incentive schemes—although cutoff-based—are predicated on inducing higher effort levels from teachers, i.e., moral hazard/hidden actions. 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. This model assumes the contract is linear, which need not be optimal. However, the solution of this model is the same as that in Hölmstrom and Milgrom (1987), which studies a static one-period model split into a number of sub-periods, where in each sub-period an agent (i.e., teacher) controls chooses the probability of success for a binomial random variable. In particular, Hölmstrom and Milgrom (1987) show that the optimal contract is linear, featuring an end-of-period payment that is a linear function of aggregated signals. The interpretation for an education context would be that, in each infinitesimal unit of time, the teacher could exert more or less effort to increase the probability a student obtains a sub-period-specific "bit" of human capital, which is measured by an end-of-year exam.

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 $\hat{q}_i = q_i + \overline{\epsilon}_i = \theta_i + \overline{\epsilon}_i = a_i + \nu_i + \overline{\epsilon}_i$. Note that the risk-neutrality of the administrator's objective implies that she can solve a separate problem for each teacher.

As Hölmstrom and Milgrom (1987) show, 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 $\nu_i + \overline{\epsilon}_i$, which we can collect 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.[17]

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

[17]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 $\overline{\epsilon}_i$ 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. This assumption is, therefore, consistent with this model's focus on a hidden action, in contrast to the hidden type specification.

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).[18] Therefore, expected output is $\mathrm{E}\left[q^*\right]=\mathrm{E}_\nu\left[a^*+\nu\right]=a^*=1/\left(\gamma(1+\xi\gamma\sigma_\eta^2)\right)$.[19] 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 $\mathrm{E}\left[q^*\right]$.

As with the hidden type model, it is important to understand how choice of estimator would affect output in this 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 ratio of signal to noise as fixed effects when class sizes are constant, meaning the contract slope would simply adjust to take into account its shrunken distribution. An implication of Proposition 6 is that we can scale the empirical Bayes estimator in Model HA to have the same variance as the fixed effects estimator. That is, we can compare estimator performance by scaling them to have the same variance and consider only the information they contain.

[18] 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.

[19] 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.

Model HA highlights the bias-variance "tradeoff" in a sense: if the variance of the fixed effects estimator increased, the resulting 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 is used in remuneration (i.e., decrease β_1). Indeed, Proposition 6 shows we can re-scale the empirical Bayes estimator when class size is constant, suggesting the use of a biased, yet lower-variance estimator could be modeled by increasing the effective error variance σ_η^2. What matters is the amount of information about the action a in the output signal (Hölmstrom (1979)).

Therefore, as with the cutoff and hidden type models, the theoretical effect of switching from empirical Bayes to fixed effects is unambiguous in the hidden action model, given the relationship between class size and teacher quality: output 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. As noted previously, Appendix E shows that the administrator with a much more general objective would also prefer fixed effects over empirical Bayes when $n(\theta)$ was negative quadratic and empirical Bayes when $n(\theta)$ was positive quadratic.

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.[20] In Section 5.1, I calibrate parameters needed to compare estimator performance in the cutoff model, which is most parsimonious. 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 back-of-the-envelope calculation of how choice of estimator would affect output in the hidden type model. Section 5.4 presents a calibration of the additional parameters of the hidden action model and computes how choice of estimator would affect output in that environment. 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.

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

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 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. This value is similar to the one used in Staiger and Rockoff (2010). 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 (2011)).[21] 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).[22] 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).[23] The average class size is 22.5 students, with a standard deviation of 5 students.

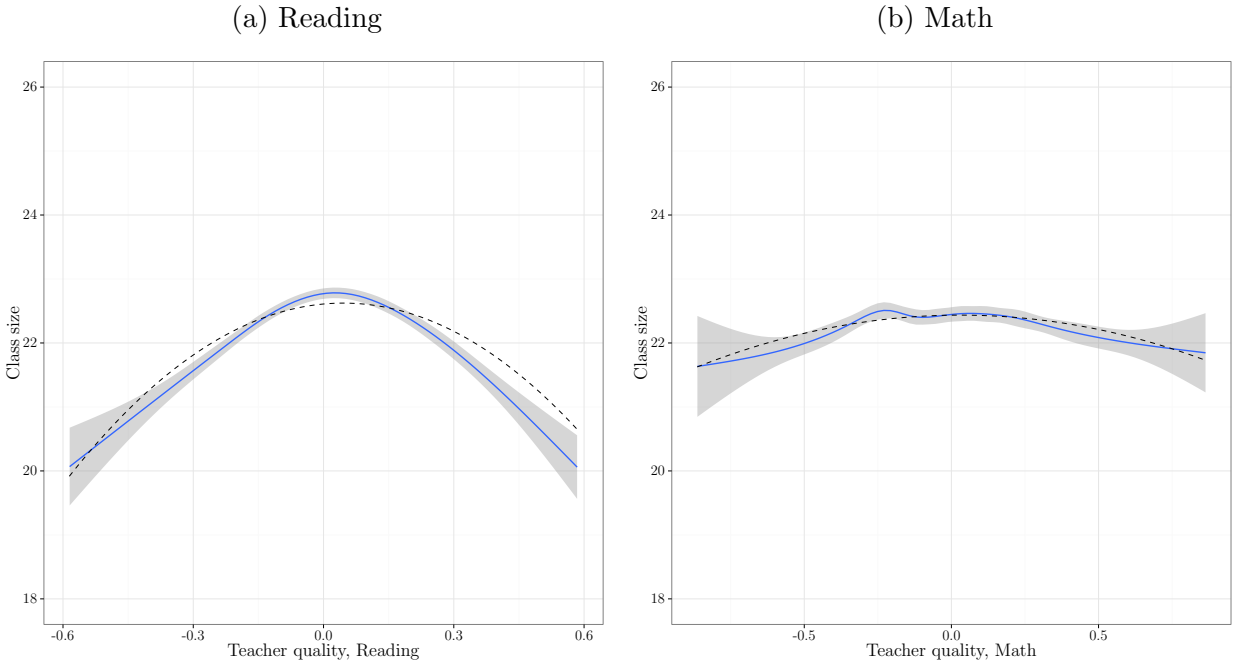
Figure 4 plots non-parametric regressions (solid blue lines) of class size on estimated teacher value-added for Reading (4a) and Math (4b). Teachers at either end of the distribution of

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

[22] The results do not appreciably change when using value-added estimates from specifications that control for subsets of these characteristics.

[23] 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.

Figure 4: The relationship between class size and teacher quality



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 4 shows the regression line fit for models in columns (1) and (3). Columns (2) and (4) are the same as regressions in (1) and (3), respectively, but exclude teachers whose estimated quality is more than two standard deviations from the population mean, 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, the observed relationship between teacher quality and class size is the worst-case scenario for the empirical Bayes estimator, as outlined by Proposition 2.

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 Table 1. The estimated relationship $\hat{n}(\hat{\theta})$ also features a mechanical negative-quadratic relationship, caused by heteroskedastic errors possible even under identically distributed class sizes. To address these issues, I calibrate $n(\theta)$ using an indirect inference approach described in Appendix D.2. Table 2 presents the calibrated

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

relationships between teacher quality and class size, $n(\theta)$, which are used for the quantitative results. 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 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.

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: Calibration details are in Appendix D.2.

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 added), I solve for the administrator's optimal cutoff policy for fixed-effects and empirical Bayes estimators, assuming a symmetric loss function.[24] 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 5a 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 line) and Math (dotted red line), for desired cutoffs ranging from the lowest to the highest teacher qualities. The right panel (5b) plots how many more expected mistakes the empirical Bayes estimator would make than the fixed effects estimator, assuming the Los Angeles school district employed 30,000 teachers.[25] 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, using fixed effects would increase the administrator's value by 2%, corresponding to the empirical Bayes estimator making 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.[26] The administrator's values from using the fixed effects and empirical Bayes estimators become comparable as the desired cutoff approaches the center of the distribution of teacher quality. In sum, 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

[24] Results are qualitatively similar under asymmetric preferences, i.e., where $\alpha \neq 1/2$; see Appendix B.3.

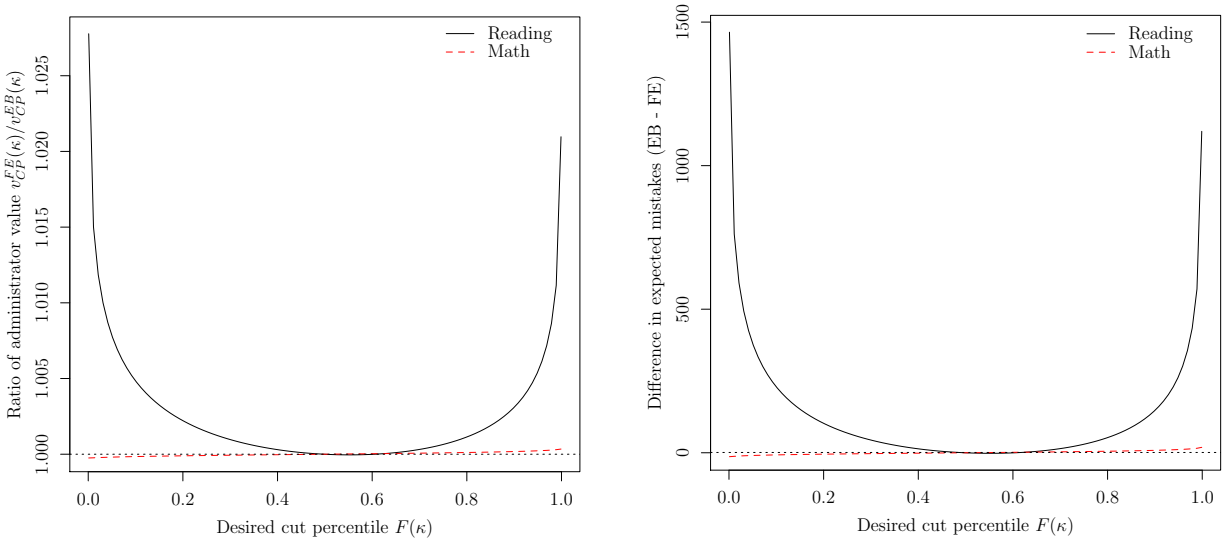
[25] The Los Angeles school district is the second-largest in the US. Though the value-added data I am using cover 30,000 teachers, more than 45,000 worked in the district in 2007 (`http://en.wikipedia.org/wiki/Los_Angeles_Unified_School_District`).

[26] 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.

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 5: 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.

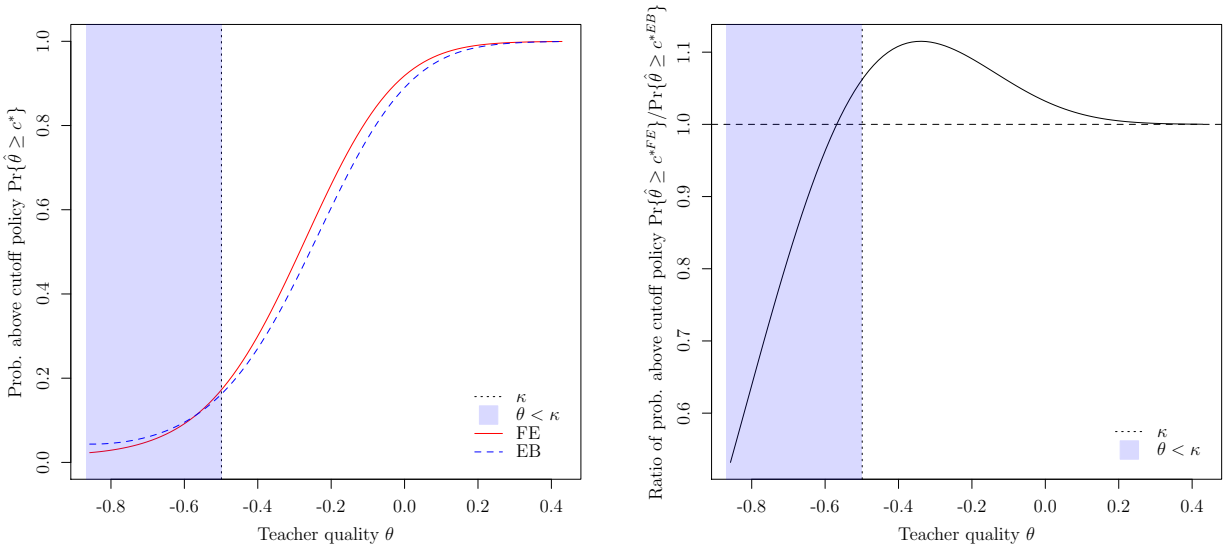
The divergence in estimator performance is largest when the desired cutoff is in the tails of true teacher quality. However, all teachers would be affected by the administrator's choice of estimator. Figure 6a plots the probability that a teacher with true quality θ, measured along the x-axis, has an 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 line) and empirical Bayes (dashed blue line) 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 6a. 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 6b, which plots the ratio of probability of the estimate being above the respective cutoff for fixed effects over empirical Bayes, i.e., $\Pr\{\hat{\theta}^{FE} \geq c^{*FE}\} / \Pr\{\hat{\theta}^{EB} \geq c^{*EB}\}$. 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. More generally, teachers over a large range of quality would be differentially affected by the estimator—that is, the impacts are not limited only to those in the extreme tails of the quality distribution.

Figure 6: 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



5.3 Quantitative Findings: Hidden Type Model

Although the cutoff-based model has an intuitive outcome space—the probability of correct classification—it would also be of interest to gauge how choice of estimator would affect output. We can also use the calibrated relationship between teacher quality and class size to form a

rough idea of how moving to an output-based retention policy would affect outcomes if we had information about the replacement cost χ.

As a rough approximation I use the two-period Model HT-2 to get an idea for how much the choice of estimator affects output. I computed output 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 then naturally viewed in terms of teacher quality.

Expected output when using empirical Bayes and the optimal reservation signal policy $\underline{q}^{*EB}(\chi, n(\theta))$, is 0.058. That is, second-period teacher quality from using empirical Bayes would be 5.8% of a standard deviation higher than it would be in a world where all teachers were retained. Expected teacher quality, and hence, output, from using fixed effects would be 0.11% higher. If instead, we used the value $\chi = 0.07$ from Staiger and Rockoff (2010), expected teacher quality in the second period would be 0.22% higher under fixed effects than when using empirical Bayes. Either way, even though they are only based on a rough example, these results suggest there is potentially a considerable benefit from using fixed effects instead of empirical Bayes to measure teacher quality.

5.4 Quantitative Illustration: Hidden Action Model

As with the hidden type environment, it would be useful to get even a rough sense of how measurement issues affect output in the real world in a hidden action environment, by using a tractable model and realistic values for model parameters, including the relationship between class size and teacher quality. The simplicity of the hidden action model affords an ancillary contribution. If the model can be calibrated then, not only can I compute how the choice of estimator would affect output, but I can also provide a rough sense of what the optimal contract *should* look like in this environment, using the calibrated parameter values. This, then, comprises an additional contribution of the current paper.

Therefore, this section takes two approaches to roughly examine how introducing output-based wages and choice of estimator, might affect educational production. 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.[27] 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 in the absence of 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 (see Appendix D.3 for details).[28] As was the case with a cutoff-based rule, the empirical Bayes estimator makes it more difficult to separate high- and low-performing teachers when the class size function is negative quadratic. This can be modeled as increasing the measurement error variance on teacher action, σ_η^2, by 3.2%.[29] The next section shows how to obtain a baseline value of σ_η^2, that is, the value of σ_η^2 under fixed effects, which is required to solve the model.

Calibration Based on Muralidharan and Sundararaman (2011) With the above assessment of how using empirical Bayes would affect measurement of output in the hidden action model, we can obtain some rough guidance from research in this area by calibrating the hidden action model using a "sophisticated" back-of-the-envelope method and data from an experimental teacher incentive pay scheme. I use the term "sophisticated" because I calibrate using equilibrium implications of the hidden action model. The objective is to use data from the study and other information as needed, to calibrate the model parameters $(\gamma, \xi, \sigma_\eta^2)$. With these in hand, then it is possible to characterize the optimal contract and the effect of using the empirical Bayes estimator on equilibrium output under this optimal contract. As I show below, values for ξ and σ_η^2 can be obtained either directly from external sources or by transforming external data. However, to calibrate the effort cost parameter γ we need to know how much teachers respond to incentive pay.

Muralidharan and Sundararaman (2011) estimate the effect of an output-based incentive scheme for teachers in the Indian state of Andhra Pradesh, in which teachers were paid according to a linear schedule, 500 rupees per percent increase in mean test scores, for test score gains

[27]The negligible relationship between teacher quality and class size for Math (see Table 2), when combined with Proposition 6, obviates having to solve the model to compare estimator performance.

[28]This exercise abstracts from the error introduced by class size uncertainty. As is shown below, this would understate the gain in output from using fixed effects instead of empirical Bayes.

[29]Of course, it would be in principle possible to also directly condition on class size. However, 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.

above 5%. The study covered two years. Of course, it would be ideal to use a study of an incentive pay program in Los Angeles, or even the US. Unfortunately, to my knowledge, none exist which provide a comparably close map to the environment developed in Model HA. Moreover, as with any mapping between theory and data, assumptions have to be made. For example, teachers are assumed to be paid a share of the amount of income generated ("output") by their action. However, we can still learn something from this exercise. First, note that CARA utility implies that risk aversion is independent of wealth, meaning the large differences in wealth between teachers in India and the US may not affect teacher actions/output (wealth differences would only affect the intercept in the contract β_0, via the outside option). Therefore, though I calibrate remaining parameters as if the setting were the teachers in Andhra Pradesh and the relationship between class size and teacher quality were the calibrated one based on LAUSD Reading data, there is reason to believe the results may also be informative about output in this paper's context of the US.[30] Second, the linear scheme employed in Muralidharan and Sundararaman (2011) allows me to cleanly map their findings to the hidden action model, as does their experimental research design, which obviates having to account for differences in output between treatment and control groups being based on selection on hidden types. Although the context is India, I convert currency into US dollars for convenience.

There were on average 3.14 teachers and 37.5 pupils per teacher in the incentive schools. Students' annual wages increased by an average of 2,156 rupees per student[31] and the average cost of the incentive scheme was 20,000 rupees.[32] Assuming that none of this amount went to administering the program and a conversion rate of 45 rupees per dollar, this corresponds to \$1,796.67 (=\$47.91×37.5) in total output produced by the average teacher and \$141.54 (=\$444.44/3.14) paid to the average teacher. Then, the slope of the contract is the per-teacher income increase (\$141.54) divided by the increase in output (\$1,796.67), or 0.0788; i.e., teachers are paid a piece rate of 7.88% of output.

There is no particular reason to assume that the incentive pay schedule in this experiment was optimal. However, note that we can exploit 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 γ. The value of γ which rationalizes this increase is then $\gamma = \beta_1/a = 0.0788/1,796.67 = 4.385 \times 10^{-5}$. Nadler and Wiswall (2011) provide evidence that teacher risk aversion matters for how incentives are structured. I set the CARA parameter 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, which is relevant for the teacher's actions, depends on the conversion from test scores to output and the variance of test scores. Suppose mean test scores $\overline{y}$

[30]Results are similar if I instead use the mean class size in the LA data.

[31]See footnote 34 on page 72 of Muralidharan and Sundararaman (2011).

[32]The incentive scheme cost an average of 10,000 rupees for each of two years.

were converted to observed output via $\hat{q} = \beta_q \overline{y}$. Then the conversion factor β_q can be calibrated by noting that the scheme increased test scores by 0.15 sd and output per teacher by \$1,796.67, giving a conversion \$11,977.78 (=\$1,796.67/0.15). The variance of test score signal can be computed by dividing the baseline variance of test score measurement error σ^2_ϵ, from Section 5.1, by the mean number of students per teacher in the data, i.e., 0.953/(37.5). To obtain the variance of income σ^2_η we then square the test-score-to-income parameter and multiply by the variance of mean test score, i.e., σ^2_η =6,076,631\$^{2}(=\$11,977.78$^2 \times$ 0.953/(37.5)).[33]

Having obtained the calibrated parameter values above, we can now solve for the optimal slope of $\beta_1^{*FE} = 0.483$ and a corresponding optimal action of $a^{*FE} = \$11,011.34$, which corresponds to an average increase in student achievement of 0.919 sd. This increase is 6 times larger than the estimated increase in student achievement stemming from the much weaker incentives provided under the experiment. In contrast, using the earlier reckoning that empirical Bayes increases the variance of η by 3.2%, using empirical Bayes would produce an optimal slope of $\beta_1^{*EB} = 0.475$ and optimal action of $a^{*EB} = \$10,832.07$, i.e., a 0.904 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 1.65% 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 effects 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 output would be affected by varying the parameters of the hidden action model. Figure 7 presents contour maps of model outcomes for a grid of points covering a wide range of alternative values of σ^2_η and ξ, ranging from one half to ten times the calibrated value of each parameter.[34] 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

[33]This is because the variance of $\hat{q}$, i.e., σ^2_η, is $\beta^2_q \sigma^2_{\overline{y}}$. Alternatively, the calibration could be in terms of test scores by scaling the CARA parameter by the increase in output per sd increase in mean test score, returning $\xi_{\overline{y}} = 6.7 \times 10^{-3} \times 11,977.78 = 80.25$. Then, using the slope of $\beta_1 = 0.0788$ and output increase of 0.15 sd, we can compute $\gamma_{\overline{y}} = 0.0788/0.15 = 0.525$. Then, setting the variance of η equal to $\sigma^2_{\eta,\overline{y}} = \sigma^2_\epsilon/37.5$ we obtain exactly the same optimal slope and action as when units are denominated in dollars.

[34]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.

does not depend on (σ_η^2, ξ) and is therefore fixed. Figure 7a is a contour map of the optimal output share when using fixed effects, or β_1^{*FE}. Figure 7b is a contour map of optimal output when using fixed effects, i.e., $\mathrm{E}\left[q^{*FE}\right]$. In both figures, the value corresponding to the calibrated values of σ_η^2 and ξ is indicated by a red dot. We can see that as teachers become more risk averse (increasing ξ) or the output measure becomes noisier (increasing σ_η^2), both incentive strength (Figure 7a) and output decrease (Figure 7b). For example, the increase in output ranges from over 3 sd in student achievement to around 0.5 sd when teachers are ten times more risk averse than their calibrated value of $\xi = 6.7e-3$; this latter figure is only about three times the estimated effect of the incentive scheme. Figure 7c is a contour map of the expected share of teacher income comprised by variable compensation when using fixed effects, i.e., $\mathrm{E}\left[\beta_1^{*FE} q^{*FE}\right] / \mathrm{E}\left[\beta_0^{*FE} + \beta_1^{*FE} q^{*FE}\right]$. As with the slope and output, this share declines as the output noise variance and degree of risk aversion increase.[35] The optimal expected share of income that is variable pay under the calibrated parameter values would be around 7%.

As interesting as these results are in their own right, the goal in this section is to quantify the difference in output stemming from using one estimator versus another. Figure 7d shows the ratio in optimal output from using fixed effects over that using empirical Bayes. We can see here that, although optimal incentive strength and output gains vary quite a bit (in ways we would naturally expect) with respect to σ_η^2 and ξ, the output gain associated with using fixed effects versus empirical Bayes ranges from a little more than 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 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 and more realistic structural model. However, the variable share of compensation in Figure 7c can provide further of guidance for, say, a reader skeptical of the calibrated values of σ_η^2 and ξ. 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; this would correspond to the upper-right triangle of Figure 7c. 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.

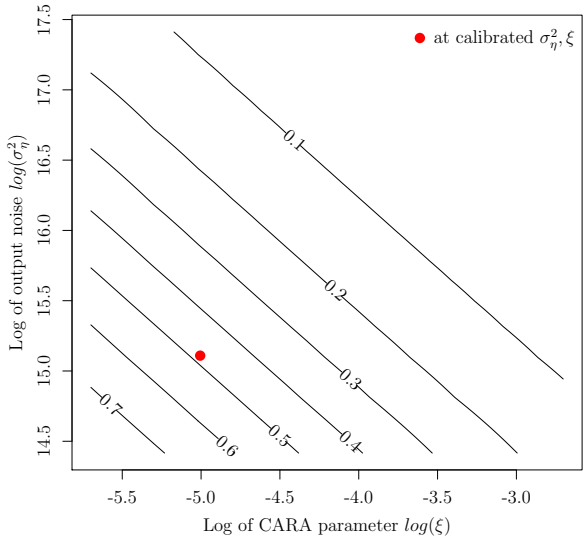
6 Discussion

While economic theory can help inform education policy, measurement issues are also important when considering how to use data in actual educational policies. Possibly because they

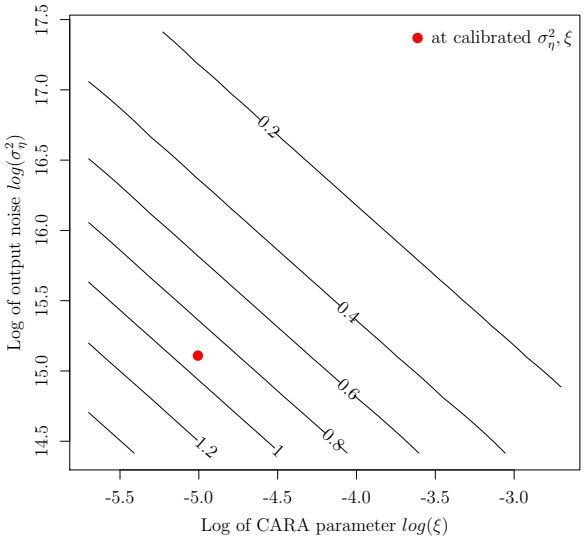
[35] This was computed using a certainty equivalent value of \$70,000, which is in the realm of teacher incomes; see Himes (2015).

Figure 7: Optimal output share and ratio of output for $(\sigma_\eta^2, \xi)-$grid

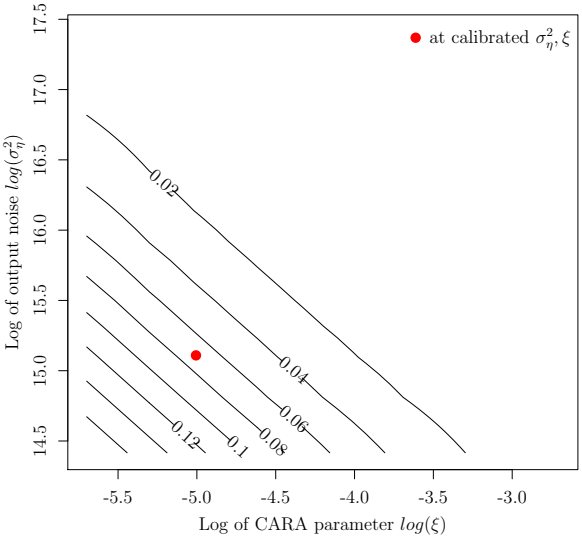
(a) Optimal output share under fixed effects, β_1^{*FE}



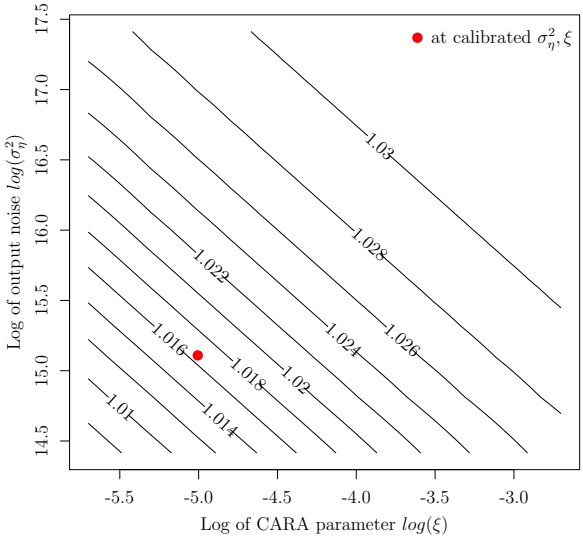
(b) Optimal output under fixed effects (sd), $\mathrm{E}\left[q^{*FE}\right]$



(c) Variable share of income, $\mathrm{E}\left[\beta_1^{*FE} q^{*FE}\right] / \mathrm{E}\left[\beta_0^{*FE} + \beta_1^{*FE} q^{*FE}\right]$



(d) Ratio of optimal output, $\mathrm{E}\left[q^{*FE}\right] / \mathrm{E}\left[q^{*EB}\right]$



minimize mean squared error, 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 even overall wages. It is not obvious this should be the case.

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 only depends on the 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, the second-largest district in the United States. At the lowest and highest percentiles of desired quality, an administrator using empirical Bayes would respectively make 10% or 7% more classification mistakes than the fixed effects estimator when classifying Los Angeles teachers based on their students' Reading test scores. Using fixed effects instead of empirical Bayes would increase output by 1.65% in the hidden action model and by between 0.11%-0.22% in the hidden type model.

Suppose an administrator had been using empirical Bayes in an incentive scheme. Would it make sense to switch to fixed effects? Of course, the relevant comparison in any economic context 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 cost of switching to fixed effects is virtually zero, or even negative, given the increased transparency of fixed effects, which may 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. Moreover, the finding of this paper—that measurement of teacher quality may affect the performance of teacher incentive schemes—could in principle be applied to findings from other work in this area, or work studying how to structure incentives and personnel decisions, based on noisy output measures.

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 the latter 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 (such as that created by the Los Angeles Times release of the value-added estimates used here). 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. Other important future work could estimate models of assignment of students to teachers.

Appendix

A Teacher Incentive Schemes

Table 3 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 3: 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.

B Cutoff Model Proofs and Extensions

B.1 Direct Conditioning on Class Size

The difference in administrator's value from using different teacher-quality estimators derives from the assumption that the administrator chooses a cutoff policy based on only test score information. Such a one-dimensional policy is quite simple and, therefore, is of considerable clear policy relevance; this demonstrated by Table 3, which documents existing incentive pay programs and shows that none condition on class size, among incentivized teachers. Additionally,

when compared with a policy that may also explicitly condition on class sizes, a test-score-based cutoff may attenuate issues of class size manipulation for the sake of affecting the administrator's posterior about the quality of a particular teacher. However, allowing the administrator to explicitly take into account class size is of theoretical interest. Therefore, this section shows how the theoretical results in Section 3 would be affected.

Now suppose the administrator, instead of only indirectly taking it into account when maximizing her utility, could instead explicitly condition on class size n_i. If n_i was a strictly monotonic function of teacher quality θ then the administrator could achieve a perfect classification of teachers by inverting $n(\theta)$—even if she ignored all teachers' test scores. A more realistic case is where there are multiple teacher qualities for at least one class size. Suppose that the distribution of teacher qualities for each class size is normally distributed. Note that, because she can explicitly condition on class size, she can hold a separate cutoff-based classification problem for each class size level; denote the administrator's value from using the fixed effects and empirical Bayes estimators as $v^{FE}_{CP,n}(\kappa)$ and $v^{EB}_{CP,n}(\kappa)$, respectively. Then by Proposition 1 the administrator would obtain the same value for either estimator given the desired cutoff κ, i.e., $v^{FE}_{CP,n}(\kappa) = v^{EB}_{CP,n}(\kappa)$ for all (n, κ). Therefore, we can without loss of generality consider only the fixed-effects estimator, with optimal cutoff policy c^{*FE}_n. Further note that the administrator's expected objective would be at least as high if she is allowed to split her original objective into one objective for each class size; if the cutoff for $c^{*FE}_{n_1} = c^{*FE}_{n_2}$ for all class sizes n_1, n_2, then her value under the separate class size scheme is the same as that from her original objective.

B.2 Administrator's Problem with Infinite Precision

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

$$u_{CP}(\theta, \hat{\theta}; c, \kappa) = \alpha 1\{\hat{\theta} \geq c \cap \theta \geq \kappa\} + (1-\alpha)1\{\hat{\theta} < c \cap \theta < \kappa\} \quad \xrightarrow{p} \alpha 1\{\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)1\{\theta < c \cap \theta < \kappa\}, \tag{11}$$

which is maximized at $c = \kappa$. The administrator's utility from using empirical Bayes estimator for the same teacher is

$$\begin{aligned} plim_{\sigma_{\bar{\epsilon}} \to 0} u_{CP}(\theta, \hat{\theta}; c, \kappa) &= \alpha 1\{\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)1\{\theta < c \cap \theta < \kappa\} \\ &= \alpha 1\{\lambda(\theta)\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)1\{\lambda(\theta)\theta < c \cap \theta < \kappa\}, \end{aligned} \tag{12}$$

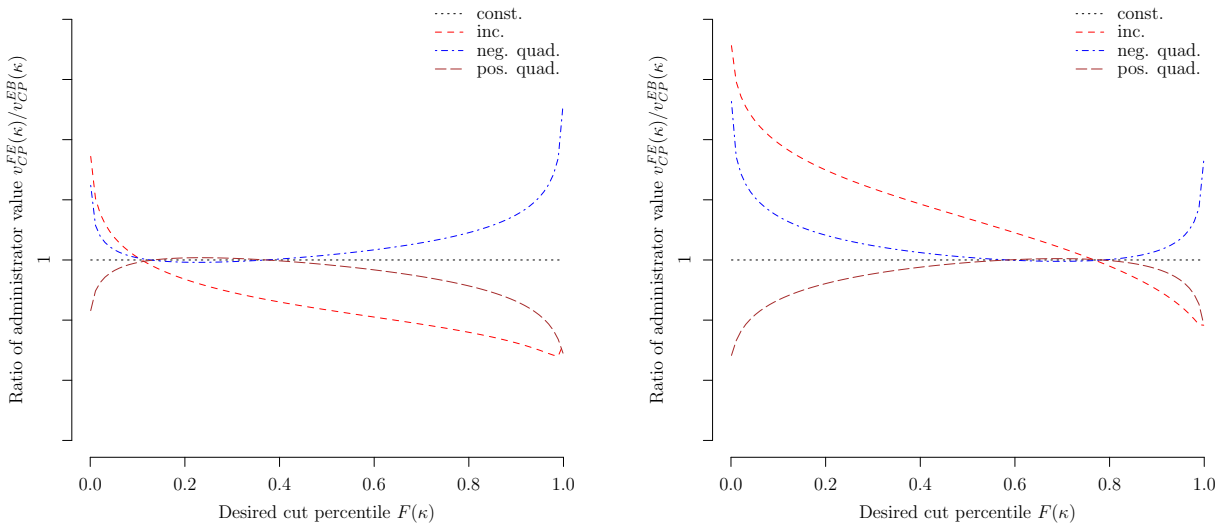
which is maximized at $c = \kappa/\lambda(F^{-1}(\kappa))$. The probabilities of the events in both (11) and (12) are all 1, giving an expected utility of 1 for all teacher qualities, which then integrates to a value of 1 for each estimator.

B.3 Asymmetric Type I and Type II Weights

The main analysis assumed the administrator equally weighed Type I and II errors, i.e., $\alpha = 1/2$. However, the administrator's preferred estimator is not sensitive to this assumption. Figure 8 plots the ratio of the administrator's value under fixed effects and empirical Bayes, by class size scenario $n(\theta)$ and desired cutoff κ, for different values of the Type I error weight. Figure 8a shows the ratio in administrator's value when $\alpha = 1/4$, or the administrator values Type I errors one-third as much as she values Type II errors. Figure 8b shows the same ratio for when $\alpha = 2/3$, i.e., the administrator values Type I errors twice as much as Type II errors. In both plots, we can see that the relative ranking of the estimators is the same as it was under the symmetric weight, $\alpha = 1/2$, scenario.

Figure 8: Difference between administrator's value under fixed effects and empirical Bayes, by class size scenario and desired cut point and weight on Type I error, α

(a) $v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa)$ when $\alpha = 1/4$ (b) $v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa)$ when $\alpha = 2/3$



B.4 Proposition 7

This section proves that fixed effects and empirical Bayes return the same value when the administrator's problem is symmetric.

Definition 1. *The administrator's problem is symmetric if $\alpha = 1/2$, $n(\theta)$ is symmetric around the population mean of teacher quality, and the administrator's desired cutoff is $\kappa = 0$.*

Proposition 7. *The administrator receives the same value from both estimators when the problem is symmetric.*

Proof. Because $n(\theta)$ is symmetric about $\theta = 0$ and $\theta_i \sim F = \mathrm{N}(0, \sigma_\theta^2)$, the distribution of θ is symmetric around its population mean of 0. The optimal c^{*EB} solves

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

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

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

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

C Extensions to Hidden Type Model HT-0

C.1 Model HT-1

Now allow $T > 2$ and let output depend on teacher experience $x_{i(j,t),t}$ according to $q_{jit} = \beta_0 + \theta_{i(j,t)} + e(x_{i(j,t),t})$, where $e(x_{it})$ represents output, net of β_0 and teacher quality, for a teacher with $t-1$ periods of prior experience.

The optimal hiring policy ψ_h is unchanged. Consider the retention decision for teachers in period $t = T$, for teachers with the same experience, $x_{it} = x_t$. Such a policy need not only apply to teachers' first years of experience; Wiswall (2013) shows that teacher quality also changes after the first few years of experience. Let $\hat{q}_{H_{it}}$ be the sample mean of teacher i's output signals realized before period t. The retention decision ψ_r still has a reservation value property, which now depends on the mean of each teacher's entire history of signals, $\hat{q}_{H_{it}}$, where the threshold now depends on the period, i.e., $\underline{q}_t = \mu - \left(\frac{\chi + e(x_t)}{\rho_t}\right)$, where $\rho_t = \sigma_\theta^2/(\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{n|H_t|})$. The reservation

signal $\underline{q}_t$ is decreasing in x_t if there are productivity gains to experience and increasing in ρ_t, due to the higher precision about teachers' true quality. Note that solution to this problem would be the same as that from HT-0, setting the replacement cost (in HT-0) to $\chi_t \equiv \chi + e(x_t)$ and using the relevant ρ_t. Also, note that considering instead periods $t < T$ would change the desired threshold quality, which could be modeled by suitably adjusting the replacement cost χ from the static model HT-0. Therefore, this sequence of per-period reservation signals can then be mapped to the cutoff-based model via a sequence of cutoff-based problems, one for each period of experience, as was done for Model HT-0. Finally, note that a similar transformation to the one above could be performed to adapt Model HT-2 (see Section C.2) to also allow for an effect of experience on output.

C.2 Model HT-2

This model augments HT-0 to allow class size to depend on teacher quality, i.e., $n_i = n(\theta)$.

As in HT-0, consider the administrator's problem in the second period. As in the cutoff model, the administrator must now integrate over the distribution of class sizes when choosing their reservation signal, meaning (8) must be adapted to obtain the administrator's value from using fixed effects:

$$v_{HT2}^{FE}(\chi) = \max_{\underline{q}^{FE}} \left(\int_{-\infty}^{\infty} \Phi\left(\frac{\underline{q}^{FE} - \theta}{\sigma_{\bar{\epsilon}}(n(\theta))} \right) dF(\theta) \right) (-\chi) + \int_{-\infty}^{\infty} \left(1 - \Phi\left(\frac{\underline{q}^{FE} - \theta}{\sigma_{\bar{\epsilon}}(n(\theta))} \right) \right) \left(\frac{\sigma_\theta^2}{\sigma_{\hat{\theta}^{FE}(n(\theta))}} \frac{\phi\left(-\underline{q}^{FE} / \sigma_{\hat{\theta}^{FE}(n(\theta))} \right)}{\Phi\left(-\underline{q}^{FE} / \sigma_{\hat{\theta}^{FE}(n(\theta))} \right)} \right) dF(\theta), \quad (13)$$

where $\sigma_{\hat{\theta}^{FE}(n(\theta))} = \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2 / n(\theta)}$ and $\sigma_{\bar{\epsilon}}(n(\theta))$ is as defined on page 11. The administrator's value from using the empirical Bayes estimator is

$$v_{HT2}^{EB}(\chi) = \max_{\underline{q}^{EB}} \left(\int_{-\infty}^{\infty} \Phi\left(\frac{\underline{q}^{EB} / \lambda(n(\theta)) - \theta}{\sigma_{\bar{\epsilon}}(n(\theta))} \right) dF(\theta) \right) (-\chi) + \int_{-\infty}^{\infty} \left(1 - \Phi\left(\frac{\underline{q}^{EB} / \lambda(n(\theta)) - \theta}{\sigma_{\bar{\epsilon}}(n(\theta))} \right) \right) \left(\frac{\sigma_\theta^2}{\sigma_{\hat{\theta}^{FE}(n(\theta))}} \frac{\phi\left(-\underline{q}^{EB} / \left(\lambda(n(\theta)) \sigma_{\hat{\theta}^{FE}(n(\theta))} \right) \right)}{\Phi\left(-\underline{q}^{EB} / \left(\lambda(n(\theta)) \sigma_{\hat{\theta}^{FE}(n(\theta))} \right) \right)} \right) dF(\theta). \quad (14)$$

Because $n(\theta)$ is no longer constant, as it was in HT-0, the reliability of signals varies by teacher and the analytical characterization of the administrator's reservation signal from Model HT-0 no longer obtains. However, as long as the MLRP holds, higher signal realizations will cause the administrator to revise her belief about teacher quality upwards, meaning Proposi-

tion 3 would still apply here. The estimator-specific reservation signals, $\underline{q}^{*FE}$ and $\underline{q}^{*EB}$, are respectively obtained by numerically solving (13) and (14).

The ranking of the administrator's utility from HT-2, by class size scenario $n(\theta)$, is the same as her ranking under the cutoff-based model. To show this, I solve for the administrator's objective for a wide range of replacement costs and under different class size scenarios: constant, negative quadratic, and positive quadratic.[36] Figure 9 shows how the relationship between class size and teacher quality would affect outcomes in HT-2. The left panel, Figure 9a, plots the ratio in her value from using the FE over the EB estimator. The constant class size scenario (dotted black line) represents a special case of HT-2 where $n(\theta) = n$. Unsurprisingly, then, we obtain the same value for all replacement costs χ, as this is simply model HT-0. Under the negative quadratic scenario (dot-dashed blue line) the administrator would obtain higher value from using fixed effects for every χ. This is exactly the same result as was obtained for different desired cutoffs (and Type I and II error weights; see Appendix B.3) under the cutoff-based model. Also, as in the cutoff-based model, the estimator ranking is reversed under the positive-quadratic class size scenario (dashed red line); i.e., she would prefer to use empirical Bayes instead of fixed effects.

The right panel, Figure 9b, plots expected output under the administrator's optimal program for each estimator and replacement cost. As expected, the difference in estimator performance when class size is not constant increases even more in replacement cost, as the output measure does not take χ into account. Intuitively, retaining teachers with true quality above a certain desired threshold—which depends on the replacement cost χ—is similar to correctly identifying teachers with true quality above a particular desired cutoff κ in the cutoff model (i.e., not making a Type I error). Unlike the cutoff model, in the hidden type models, the administrator faces the same (replacement) cost for obtaining new teachers; that is, the cost portion of her objective does not directly depend on teachers' true quality θ.

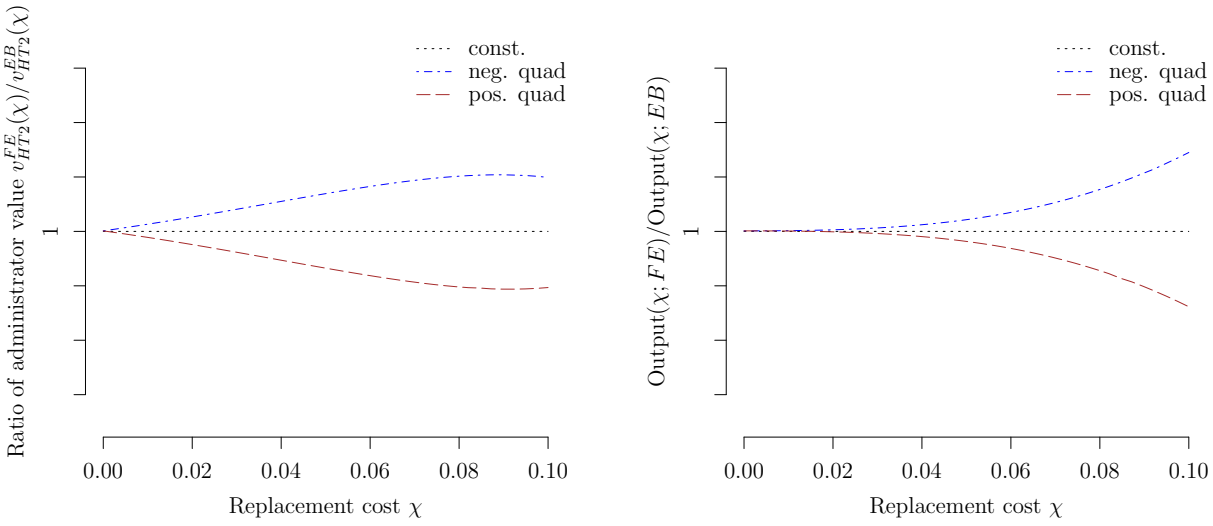
As with model HT-1, an environment with multiple periods could be modeled by suitably adjusting the desired threshold quality. For example, adding more periods could be accommodated by decreasing the replacement cost, as the administrator would have a relatively higher gain from replacing when there are more periods of output. Because they range from a cost of zero to twice the estimated difference in value added between a teacher with five years experience and no experience, the calculations presented in Figure 9 then likely encompass costs for multi-period environments as well.

The takeaway from this section is that (i) the administrator's preferred estimator depends on the class size scenario $n(\theta)$, (ii) though the difference in values from using either estimator

[36] 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; this implies a 0.25 sd difference acquired in the first five years of experience. Therefore, I set $\chi = 0.25\sigma_\theta = 0.054$ and then use a range for the replacement cost running from zero to 0.10, approximately twice this value.

Figure 9: Difference between administrator's value under fixed effects and empirical Bayes, by class size scenario and replacement cost χ

(a) $v^{FE}_{HT2}(\chi)/v^{EB}_{HT2}(\chi)$ (b) Ratio of period-2 output, FE/EB



depends on other model parameters (T, χ), the preferred estimator does not, and (iii) the administrator would prefer the same estimator in HT-2 as she would in the cutoff model.

D Details for Quantitative Exercises

D.1 Calibrated Error Variances

I calibrate σ^2_θ and σ^2_ϵ from Table B-2 of Schochet and Chiang (2012) normalizing the total variance to one. To most closely match a policy where an administrator would like to rank teachers across a school district, I calibrate $\sigma^2_\theta = 0.046$ by summing the average of school- and teacher-level variances in random effects. To most closely approximate an environment where both student and aggregate-level shocks may affect student test scores, I calibrate $\sigma^2_\epsilon = 0.953$ by summing the average of class- and student-level variances in random effects. Note that, due to the much greater student-level error variance, the approximate sizes of σ^2_θ and σ^2_ϵ are approximately the same if school-level variances are excluded from σ^2_θ or class-level variances are excluded from σ^2_ϵ, lending robustness to the quantitative findings.

D.2 Heteroskedasticity Correction for Relationship Between Class Size and Teacher Quality

The advantage of the indirect inference approach is that it can be implemented using a vector of auxiliary moments which do not necessarily correspond to structural econometric parameters. This is useful in the current context where the micro-data to directly correct for heteroskedasticity are not available.[37]

Indirect Inference Algorithm The following is done separately for Reading and Math.

0. Estimate the relationship between class size (n_i) and teacher i's estimated quality in the subject ($\hat{\theta}_i$) by running the regression $n_i = \beta_0^{data} + \beta_1^{data}\hat{\theta}_i + \beta_2^{data}(\hat{\theta}_i)^2 + e_i$. The regression coefficients $\left(\hat{\beta}_0^{data}, \hat{\beta}_1^{data}, \hat{\beta}_2^{data}\right)$ and residual standard error $\hat{\sigma}_e^{data}$ form the first set auxiliary parameters to fit. Compute the 25th, 50th, and 75th percentiles of the empirical distribution of class sizes, $(n_{p25}^{data}, n_{p50}^{data}, n_{p75}^{data})$. These are the remaining auxiliary parameters. The target vector of auxiliary parameters is then $\left(\hat{\beta}_0^{data}, \hat{\beta}_1^{data}, \hat{\beta}_2^{data}, \hat{\sigma_e}^{data}, n_{p25}^{data}, n_{p50}^{data}, n_{p75}^{data}\right)$.

1. Given σ_θ^2, simulate teacher quality θ_i^{sim} once for each teacher in the sample. (Recall the population mean has been normalized to 0).

2. Simulate the random component of class sizes $n_{i,i.i.d}^{sim}$, which is distributed normal with mean zero and standard deviation $\sigma_{n_{sim}}$. As described below, this algorithm chooses the parameter $\sigma_{n_{sim}}$. Note these are independent from teacher quality to get an idea of the role heteroskedasticity plays.

3. Assign *incremental class sizes* according to $n^{inc}(\theta_i^{sim}) = a_0 + a_1\theta_i^{sim} + a_2(\theta_i^{sim})^2$. As described below, this algorithm chooses the parameters (a_0, a_1, a_2). The final simulated class size for teacher i is then $n_i^{sim} = \text{round}\{n_{i,i.i.d}^{sim} + n^{inc}(\theta_i^{sim})\}$, i.e., class sizes are integer-valued.

4. Given σ_ϵ^2 and n_i^{sim} simulate an average shock for each teacher, $\bar{\epsilon}_i^{sim}$; form simulated estimated teacher quality according to $\hat{\theta}_i^{sim} = \theta_i^{sim} + \bar{\epsilon}_i^{sim}$.

5. Regress $n_i^{sim} = \beta_0^{sim} + \beta_1^{sim}\hat{\theta}_i^{sim} + \beta_2^{sim}(\hat{\theta}_i^{sim})^2$, estimating the auxiliary coefficients $(\hat{\beta}_0^{sim}, \hat{\beta}_1^{sim}, \hat{\beta}_2^{sim})$ and auxiliary residual standard error $\hat{\sigma}_e^{sim}$. Compute the 25th, 50th, and 75th percentiles of the simulated distribution of class sizes, $(n_{p25}^{sim}, n_{p50}^{sim}, n_{p75}^{sim})$. The simulated vector of auxiliary parameters is then $\left(\hat{\beta}_0^{sim}, \hat{\beta}_1^{sim}, \hat{\beta}_2^{sim}, \hat{\sigma}_e^{sim}, n_{p25}^{sim}, n_{p50}^{sim}, n_{p75}^{sim}\right)$.

[37] If micro-data had been available, then one could in principle use an approach like the one in Lockwood and McCaffrey (2014) to account for the nonlinearities produced by heteroskedastic errors.

6. Compute the Euclidean distance between target auxiliary parameters and simulated auxiliary parameters (e.g., $\hat{\beta}_0^{data}$ and $\hat{\beta}_0^{sim}$, respectively) as a function of the parameters governing class size, $d(a_0, a_1, a_2, \sigma_{n_{sim}})$.

Repeat steps 1-6 for the vector $(a_0, a_1, a_2, \sigma_{n_{sim}})$ until the distance between data and simulated auxiliary moments is minimized.

D.3 Details for Quantitative Illustration for Hidden Action Model

Output in the hidden action model depends on several parameters, including the variance of measurement error on output, σ_η^2. I adjust the error variance in several steps, using Reading test scores as the measure:

1. Simulate teacher quality, class sizes, and measurement errors using the parameters from Section 5.1, for 30,000 teachers. Each simulated teacher then has a simulated quality θ_i^s and a simulated fixed-effect estimate $\hat{\theta}_i^{s,FE}$.

2. Use the empirical Bayes weights $\lambda(\cdot)$ to generate simulated EB measures of teacher quality according to $\hat{\theta}_i^{s,EB} = \lambda(n(\theta_i^s))\hat{\theta}_i^{s,FE}$.

3. Standardize θ_i^s, $\hat{\theta}_i^{s,FE}$, and $\hat{\theta}_i^{s,EB}$ to have variances of 1, to make the residual variances comparable.

4. Finally, I estimate the residual variance from a regression of standardized $\hat{\theta}_i^{s,FE}$ on the standardized true (simulated) quality θ_i^s and the residual variance from a regression of standardized empirical Bayes measure $\hat{\theta}_i^{s,EB}$ on standardized true (simulated) quality. The ratio of residual variances, or amount unexplained in each regression, tells us how much more (or less) the fixed effects estimator would inform the administrator about teacher quality.

The regression results, shown in Table 4, indicate that the fixed-effects estimator explains about 3.2% more variation in teacher quality than the empirical Bayes estimator ($1-0.6956^2/0.7070^2 = 0.032$). That is, the fact that the EB estimator makes it more difficult to separate high- and low-performing teachers when the class size function is negative quadratic, as it is in the data, can be modeled as increasing the measurement error variance on teacher output, σ_η^2, by this amount.

Table 4: Regressions of simulated teacher quality on FE and EB estimates

	Dependent variable:	
	θ^s (standardized)	
	(1)	(2)
$\hat{\theta}^{s,FE}$ (standardized)	0.718^{***}	
	(0.004)	
$\hat{\theta}^{s,EB}$ (standardized)		0.707^{***}
		(0.004)
Constant	0.002	-0.001
	(0.004)	(0.004)
Observations	30,000	30,000
R^2	0.516	0.500
Residual Std. Error (df = 29998)	0.696	0.707

Note: *p<0.1; **p<0.05; ***p<0.01

E Alternative Specification of Administrator's Objective

Characterizing optimal contracts is difficult to do in a general setting. Therefore, in this section I take a different approach and instead examine a problem qualitatively consistent with economic environments of interest. Specifically, I consider an alternative objective in which the administrator's value is increasing in the product of teacher quality and reward assigned to that teacher. For example, the administrator may want to assign higher wages to, or only retain, higher-quality teachers. This objective is reasonable so long as an increase in estimated quality does not decrease the reward assigned to a teacher and the administrator would like for higher-quality teachers to receive rewards no lower than those assigned to lower-quality teachers.

I first proceed by considering linear incentive schemes, which have been extensively studied in the literature. Next, I show how the same result holds when incentives are weakly increasing in estimated teacher quality, so long as the administrator's objective has the intuitive property of being weakly increasing in the product of the incentive and true teacher quality. For example, this specification is consistent with the cutoff and hidden type models.

Consider the following objective for the administrator from assigning a reward $w(\hat{\theta})$ to

teacher with true quality θ and quality signal $\hat{\theta}$:

$$u_{mono}(\hat{\theta}, \theta; w(\cdot)) = w(\hat{\theta})\theta, \tag{15}$$

where $\hat{\theta}_2 > \hat{\theta}_1$ implies $w(\hat{\theta}_2) \geq w(\hat{\theta}_1)$, i.e., the "reward"—which could represent monetary compensation such as bonuses or raises, or a retention decision—is weakly increasing in estimated quality. The administrator's value is obtained by integrating (15) over the distribution of teacher quality:

$$v_{mono}(\lambda(\cdot)) = \int_{\theta} \left(\int_{\hat{\theta}} w^*(\hat{\theta}) dG_{\hat{\theta}}(\hat{\theta}|\theta; \lambda(\cdot)) \right) \theta dF(\theta), \tag{16}$$

where $w^*(\cdot)$ is the optimal reward schedule and $G_{\hat{\theta}}(\cdot)$ is the distribution quality signal $\hat{\theta}$, which depends on true quality and the relationship between class size and teacher quality $\lambda(\cdot)$. Although the specification of the administrator's objective in (15) may not necessarily correspond to any economic environment, results pertaining to the estimator rankings are unchanged when we instead study the following objective:

$$u_{mono,g}(\hat{\theta}, \theta; w(\cdot)) = u_w(w(\hat{\theta}))u_\theta(\theta), \tag{17}$$

where u_w and u_θ are both weakly increasing. For example, the administrator wanting to retain higher-quality teachers or give them bonus payments, or having a wage schedule increasing in estimated quality are both consistent with (17). What matters is that, by providing the correct incentives, the administrator assigns higher rewards (or lower punishments) when she sees a higher signal of teacher quality. This is a very natural assumption.

Linear Reward First, suppose $w^*(\hat{\theta}) = \beta_0^* + \beta_1^* \hat{\theta}$, i.e., the optimal reward schedule is a linear function of estimated teacher quality. Note that, although this reward function is the same as in Model HA, the case here does not necessarily assume the underlying unobserved input is a choice of action by the teacher. This type of contract has been studied extensively in the literature, perhaps due to its tractability and simplicity (see, e.g., Tincani (2012), Behrman et al. (2016), Rothstein (2014)). Substituting in for estimated teacher quality, we get the administrator's optimized objective for a teacher with true quality θ_i:

$$\begin{aligned}
\mathrm{E}\left[u_{mono}(\hat{\theta}, \theta; w^*(\cdot)\right] &= \mathrm{E}\left[\theta_i(\beta_0^* + \beta_1^* \hat{\theta}_i)\right] \\
&= \mathrm{E}\left[\theta_i \beta_0^* + \beta_1^* \theta_i \lambda(\theta_i)\left(\theta_i + \bar{\epsilon}_i\right)\right] \\
&= \mathrm{E}\left[\theta_i \beta_0^* + \beta_1^* \lambda(\theta_i)\theta_i^2 + \beta_1^* \lambda(\theta_i)\theta_i \bar{\epsilon}_i\right] \\
&= \theta_i \beta_0^* + \beta_1^* \lambda(\theta_i)\theta_i^2,
\end{aligned} \tag{18}$$

where we go from the third to fourth line because $\mathrm{E}\left[\overline{\epsilon}_i|\theta_i\right] = 0$. By the envelope theorem, the optimal reward parameters (β_0^*, β_1^*) would not vary with an infinitesimal change in $\lambda(\cdot)$. We can see that a negative-quadratic relationship of class size (and subsequently, $\lambda(\theta_i)$) and teacher quality θ reduces the above amount, e.g., for a high- or low-quality teacher. Integrating over all teacher qualities, this means the administrator would prefer to use the fixed effects estimator for this type of incentive scheme. If class size is constant in teacher quality, then by adjusting (β_0^*, β_1^*), the administrator could obtain the same expected objective by using either the fixed effect or empirical Bayes estimator of teacher quality; this result is identical to that also derived for the cutoff, hidden type, and hidden action models.

Increasing Reward However, without further structure, there is no theoretical reason to assume the optimal incentive scheme would be linear in estimated teacher quality. Therefore, I next consider the case where an administrator may employ any scheme that is weakly increasing in estimated quality.

According to (15), the administrator's optimized objective for a teacher with true quality θ_i and shock $\overline{\epsilon}_i$ is

$$\begin{aligned} u_{mono}(\theta_i, \hat{\theta}_i; w^*(\cdot)) &= w^*(\hat{\theta}_i)\theta_i = w^*(\lambda(\theta_i)\hat{\theta}_i^{FE})\theta_i = w^*(\lambda(\theta_i)\left(\theta_i + \epsilon_i\right))\theta_i \\ &= w^*(\lambda(\theta_i)\theta_i + \lambda(\theta_i)\epsilon_i)\theta_i, \end{aligned} \tag{19}$$

where the potential nonlinearity of $w^*(\cdot)$ means we cannot remove the $\overline{\epsilon}_i$ from inside the wage function as we did when the wage was linear. The closer $w^*(\cdot)$ is to linear, the closer is the following approximation:

$$\mathrm{E}\left[u_{mono}(\theta_i, \hat{\theta}_i; w^*(\cdot))\right] \approx w^*(\lambda(\theta_i)\theta_i)\theta_i. \tag{20}$$

By examining (20), we can see that, intuitively, if we consider teachers with positive true quality, the expression will be higher when we change infinitesimally $\lambda(\cdot)$ to be more increasing in θ. Conversely, the expression will be higher when $\lambda(\cdot)$ is infinitesimally changed to be more decreasing in θ for teachers with negative true quality. Due to the infinitesimal nature of these changes to $\lambda(\cdot)$ and the prior optimality of $w^*(\cdot)$, we can apply the envelope theorem and disregard the effect of changes to $w^*(\cdot)$ on the administrator's expected objective. Put together, a positive-quadratic relationship between class size and teacher quality would increase the administrator's expected utility from using the empirical Bayes estimator of teacher quality, while a negative-quadratic relationship would decrease the administrator's utility from using empirical Bayes.

However, (20) makes the unattractive assumption that $w^*(\cdot)$ is "close to" linear. To more

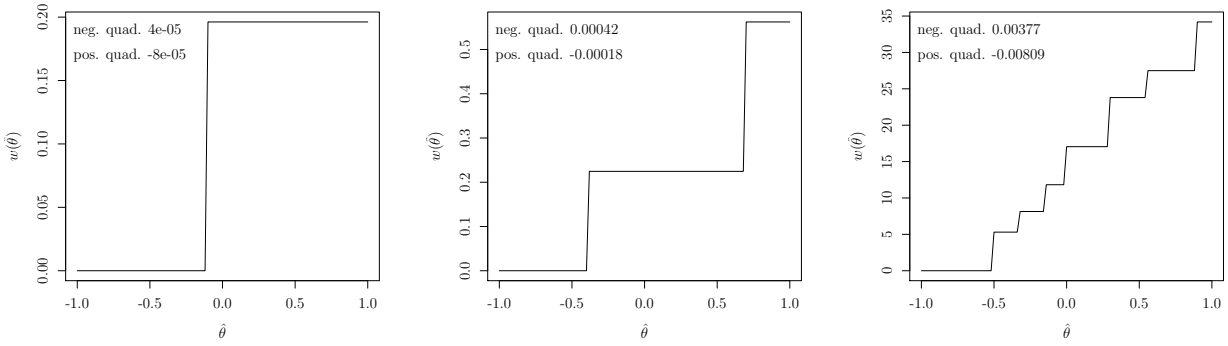
generally verify that the administrator's objective (15) is higher under fixed effects when class sizes are negative quadratic and higher under empirical Bayes when class sizes are positive quadratic, I randomly generated 100 nondecreasing wage schedules. For each schedule, I computed the derivative of (16) with respect to a convexifying parameter δ, which represents the "share" of the estimator $\hat{\theta}$ represented by the fixed effects estimator used in evaluating (16). That is, $\hat{\theta}_i = \delta\hat{\theta}_i^{FE} + (1-\delta)\lambda_i\hat{\theta}_i^{FE}$. Increasing δ puts more weight on fixed effects, and, therefore, less weight on empirical Bayes and $\lambda(\cdot)$; where $\lambda(\cdot)$ is adjusted to equalize the mean weight, integrating over the joint distribution of teacher quality and measurement error, under the initial share δ and perturbed one.

Figure 10 presents six randomly generated reward functions and the associated derivative with respect to δ under the negative-quadratic and positive-quadratic class size scenarios.[38] For example, Figure 10a presents an optimal reward function $w^*(\cdot)$ that is essentially a cutoff rule assigning a bonus (or retaining) teachers above a certain value in the estimated quality distribution $\hat{\theta}$, showing how the objective (15) can capture meaningful economic environments, such as the cutoff model and hidden type model. We can see at the top left that the derivative with respect to share is 4×10^{-5} when class size is negative quadratic in teacher quality. This positive value means the administrator would prefer to increase the weight on fixed effects in this case. We can also see at the top left that the derivative with respect to share is -8×10^{-5} when class size is positive quadratic in teacher quality, i.e., the administrator would prefer to decrease the weight on fixed effects in this case, as expected. The reward function in Figure 10f, on the other hand, has a flat spot below which the reward is flat—say, where teachers are not retained—and then a reward that is roughly increasing linearly in measured quality, similar to that in (18). This pattern holds for the rich variety of simulated wage functions, so long as monotonicity of $w^*(\cdot)$ is not violated. On average, the derivative of the administrator's objective (16) is 9.13×10^{-3} and -7.81×10^{-3} when class size is respectively negative- and positive-quadratic in teacher quality. That is, fixed effects would be preferred by the administrator when class size is negative quadratic in teacher quality and empirical Bayes would be preferred in the opposite scenario.

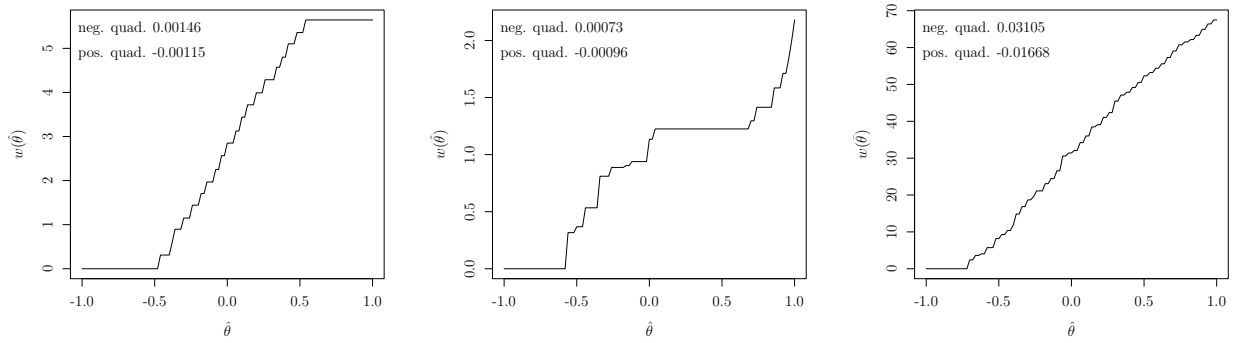
[38] Details of the simulation algorithm are available upon request.

Figure 10: Reward functions $w^*(\cdot)$ and derivative w.r.t. share δ for selected simulations, by $n(\theta)$

(a) Simulation 46 (b) Simulation 14 (c) Simulation 23



(d) Simulation 35 (e) Simulation 9 (f) Simulation 73



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.

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.

Cawley, J., J. Heckman and E. Vytlacil, "On Policies to Reward the Value Added by Educators," *Review of Economics and Statistics*, 81(4):720–727, 1999.

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.

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.

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.

Glazerman, S., H. Chiang, A. Wellington, J. Constantine and D. Player, "Impacts of Performance Pay Under the Teacher Incentive Fund: Study Design Report." *Mathematica Policy Research, Inc.*, 2011.

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.

Goldhaber, D. D. and D. J. Brewer, "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables On Educational Productivity." *Journal of Human Resources*, 32(3), 1997.

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

Green, J. and N. Stokey, "A Comparison of Tournaments and Contracts," *Journal of Political Economy*, 91(3):349–364, 1983.

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.

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.

Himes, T., "LAUSD Educators Typically Earned $75,504 Last Year," *Los Angeles Daily News*, 2015.

Hölmstrom, B., "Moral Hazard and Observability," *The 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.

—, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *JL Econ. & Org.*, 1991.

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.

Jacob, B. A. and L. Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics*, 26(1):101–136, 2008.

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.

Kane, T. J., D. F. McCaffrey, T. Miller and D. O. Staiger, "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment," *Research Paper. MET Project. Bill & Melinda Gates Foundation*, 2013.

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.

Kane, T. J. and D. O. Staiger, "Estimating Teacher Impacts On Student Achievement: An Experimental Evaluation," *NBER Working Paper*, 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.

Lang, K., "Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member," *Journal of Economic Perspectives*, 24(3):167–181, 2010.

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

Lazear, E. and S. Rosen, "Rank-Order Tournaments As Optimum Labor Contracts," *The Journal of Political Economy*, 1981.

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

Lockwood, J. and D. McCaffrey, "Should Nonlinear Functions of Test Scores be Used as Covariates in a Regression Model?" in R. Lissitz and H. Jiang, eds., "Value-added Modeling and Growth Modeling with Particular Application to Teacher and School Effectiveness," chap. 1, pp. 1–36, Information Age, Charlotte, NC, 2014.

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.

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

Nadler, C. and M. Wiswall, "Risk Aversion and Support for Merit Pay: Theory and Evidence From Minnesota's Q Comp Program," *Education Finance and Policy*, 6(1):75–104, 2011.

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

Podgursky, M. and M. Springer, "Teacher Performance Pay: A Review," *National Center on Performance Incentives*, pp. 2006–01, 2006.

—, "Teacher Compensation Systems in the United States K-12 Public School System," *National Tax Journal*, 64(1):165–192, 2011.

Rivkin, S., E. Hanushek and J. 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.

Rothschild, M., "Searching for the Lowest Price When the Distribution of Prices Is Unknown," *Journal of Political Economy*, 82(4):689–711, 1974.

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," *The Quarterly Journal of Economics*, 125(1):175–214, 2010.

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

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.

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.

Teixeira-Pinto, A. and S.-L. T. Normand, "Correlated Bivariate Continuous and Binary Outcomes: Issues and Applications," *Statistics in Medicine*, 28(13):1753–1773, 2009.

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

Todd, P. E. and K. I. Wolpin, "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The 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*, 2013.