Mehta, Nirav

Working Paper

Measuring quality for use in incentive schemes: The case of "shrinkage" estimators

CHCP Working Paper, No. 2017-25

Provided in Cooperation with:
Centre for Human Capital & Productivity (CHCP), Department of Economics, University of Western Ontario

Suggested Citation: Mehta, Nirav (2017) : Measuring quality for use in incentive schemes: The case of "shrinkage" estimators, CHCP Working Paper, No. 2017-25, The University of Western Ontario, Centre for Human Capital and Productivity (CHCP), London (Ontario)

This Version is available at:
https://hdl.handle.net/10419/180870

2017

2017-25 Measuring Quality for Use in Incentive Schemes: The Case of "Shrinkage" Estimators

Nirav Mehta

Measuring Quality for Use in Incentive Schemes: The Case of "Shrinkage" Estimators

by

Nirav Mehta

Working Paper # 2017-25 **August 2017**

Centre for Human Capital and Productivity (CHCP)

Working Paper Series

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

Measuring Quality for Use in Incentive Schemes:

The Case of "Shrinkage" Estimators

Nirav Mehta
University of Western Ontario *

August 16, 2017

Abstract

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

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

JEL codes: J01, I21, I28, D81

*nirav.mehta@uwo.ca

1 Introduction

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

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

This paper examines whether a utility-maximizing decisionmaker in an asymmetric information environment could improve upon the shrinkage estimators pervasive in research and practice. Because real-world incentive schemes are typically quite simple in structure (Stiglitz (1991); Ferrall and Shearer (1999)), I first examine a practical and concrete cutoff model, where the decisionmaker classifies agents with respect to a desired threshold to minimize a weighted sum of the expected Type I and Type II errors. To study how measurement affects output, I next study hidden type (adverse selection) and hidden action (moral hazard) models, where agent quality is, respectively, fixed and endogenous. Optimal policy in the hidden type model is a stopping rule, or reservation quality measure, which suggests an economic intuition for the cutoff model. Optimal policy in the hidden action model, which is based on Hölmstrom and Milgrom (1987), is linear in the quality measure.

In each environment, the decisionmaker chooses the best estimator from a set containing

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

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

I develop the analysis using a highly policy-relevant application: the measurement of teacher quality. The fact that only a small amount of variation in student achievement is explained by teachers' observed characteristics (Hanushek (1986); Rivkin et al. (2005)) and evidence that teacher quality is an important determinant of human capital (Hanushek (2011); Chetty et al. (2014a)) have spurred the introduction of teacher incentive schemes. For example, President Obama's Race to the Top initiative incentivizes states to adopt incentive pay schemes and the Teacher Advancement Program has introduced performance-based bonuses to over 20,000 teachers serving over 200,000 students across the U.S.[1] In addition to being a linchpin of education reform, teacher incentive schemes may be the most visible incarnation of performance-based incentives in the public sector. The concern that teacher quality measures can be quite noisy (Baker and Barton (2010)), which may subject teachers to undue risk if directly used in high-powered incentive schemes, and the fact that they minimize mean squared error has motivated the use of shrinkage estimators—most commonly "empirical Bayes"—in this application.[2]

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

[2] For example, American Federation of Teachers President Randi Weingarten said in a 2012 interview about releasing VA scores to the public: "I fought against it because we knew value-added was based on a series of assumptions and not ready for prime-time. But back then, we didn't realize the error rates could be as high as 50 percent!" (Goldstein (2012)).

The theoretical environments described above are relevant for studying teacher quality, where the decisionmaker could be a school district administrator, a public official interested in cost-effective ways of increasing educational production. As I document in Appendix A, the cutoff model matches the structure of the vast majority of existing teacher incentive schemes, which typically use empirical Bayes to measure teacher quality when assigning bonuses or even dismissing teachers; it could also be used to model pay-for-percentile-type schemes, which are tournament-based schemes that have recently become popular in education policy debates (Barlevy and Neal (2012)).[3] Optimal policies in the asymmetric information models take on the natural interpretation of reward, or wage, schedules that weakly increase in measured quality. The optimal stopping rule in the hidden type model is a step function, where teachers with above-threshold measures receive the same (positive) salary and those below receive a wage of zero (meaning they are dismissed). The optimal wage schedule in the hidden action model is a constant base salary, with a performance-based bonus linear (and increasing) in measured quality.

The aforementioned econometric endogeneity concern is germane to measuring teacher quality. The idea that class size can reflect information about teacher quality has theoretical precedent (Lazear (2001); Barrett and Toma (2013)) and researchers have found that higher-quality teachers tend to have more favorable working conditions, in terms of student characteristics (Player (2010), Clotfelter et al. (2006)); it is not a big leap to extend this reasoning to class size. Therefore, an estimator that minimizes mean squared error may not maximize the administrator's value, e.g., if shrinking makes it harder to reward high effort or fire the worst teachers.

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

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

of the distribution to retain high-quality teachers. This paper remains agnostic about the source of this relationship; all that matters for the study of estimator performance is that such a relationship exists. Examining the source of the relationship between class size and teacher quality would be interesting for future research.

Finally, I calibrate additional parameters to quantitatively compare the prospective performance of the estimators, finding nontrivial benefits to using raw quality measures. For example, in the cutoff model an administrator would make 9% more classification errors when using empirical Bayes when seeking to identify teachers in the bottom 1% in Reading value-added and switching from empirical Bayes would increase output by 2% in the hidden action model. The performance of the estimators in the cutoff model differs most at the tails of the distribution of teacher quality, which is important for identifying either high- or low-quality teachers, the focus of existing teacher incentive schemes (e.g., the Washington D.C. Schools Chancellor fired 241 teachers in 2010 based on performance measures (Turque (2010)).[4] The sheer number of schemes and affected teachers and students and increasing policy support for teacher incentive schemes point to substantial gains from using the preferred estimator for the relevant context, especially in light of the relatively costless "intervention" of adopting a simpler quality measure.

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

2 Statistical Background

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

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

increase their use in research and policy.

Manski (2004) writes that "statisticians studying estimation have long made progress by restricting attention to tractable classes of estimators; for example, linear unbiased or asymptotic normal ones," (page 1231). In the same vein, I consider a set of estimators containing a (perhaps) naive estimator based on the "raw" data, which in a value-added framework would correspond to unbiased fixed effects, and the ubiquitous mean-square-minimizing empirical Bayes. I show below in Remark 1 that this is a natural set to consider.[5] Clearly, the *unconstrained* class of optimal estimators would potentially condition on all available information, because the administrator could simply ignore information that was not valuable (Hölmstrom (1979)). No existing teacher incentive scheme does this. Therefore, in each environment I focus on the less trivial and more relevant case where the administrator chooses the (constrained) optimal estimator from the set described above.

Teacher quality is distributed according to $\theta_i \sim F = \mathrm{N}(0, \sigma_\theta^2)$, where F is known.[6] As discussed in the introduction, the number of students assigned to teacher i, n_i, may depend on i's quality. For simplicity, I assume that class size depends on θ, where I sometimes denote this dependence by writing $n(\theta)$.[7] 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 adopt this spare technology to simplify 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 distribution is also normal, giving $\hat{\theta}_i^{EB} = \lambda_i \hat{\theta}_i^{FE} + (1 - \lambda_i) \underbrace{\mathrm{E}[\theta]}_{0} = \lambda_i \hat{\theta}_i^{FE} =$

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

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

[7] 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

$\lambda_i(\theta_i + \bar{\epsilon}_i)$, 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 (signal) relative to the estimated variation using the fixed effects estimator (signal plus noise).[8] 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.[9] Note the empirical Bayes estimate for a particular teacher's quality is biased, i.e., $\mathrm{E}_{\bar{\epsilon}}\left[\hat{\theta}_i^{EB}\right] = \lambda(\theta)\theta_i \neq \theta_i$, but also has a lower variance. Though the exposition here is for fixed effects and empirical Bayes estimators, this bias-variance tradeoff would also apply to comparisons of other shrunken versus unshrunken estimators.

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

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

3 Cutoff-Based Model

In this section I develop a cutoff model, which formalizes the objective of utility-maximizing decisionmaker, 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

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

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

expected objective over all teachers in the district.[10]

I begin with this model for several reasons. First, as will be shown below, her objective can be measured in terms of the number of correct and incorrect classifications with respect to the desired cutoff, embedding the administrator's objective in a natural metric: the expected number of mistakes. Second, a discrete policy is a natural fit for modeling discrete real-world policies like retention, making the analysis in this paper highly relevant for the most pervasive, and perhaps the most contentious, education policy debates.[11] Third, even though they are not obliged to take such a form, almost all existing teacher incentive schemes for public school teachers are cutoff-based, making this model's results immediately applicable to the vast majority of existing teacher incentive schemes; as noted by Stiglitz (1991) and Ferrall and Shearer (1999), real-world incentive schemes typically take very simple forms. Fourth, related literature also considers cutoff-based policies, e.g., Staiger and Rockoff (2010), Hanushek (2011), Tincani (2012), Chetty et al. (2014b), and Rothstein (2014). Finally, the cutoff-based model's flexibility allows us to be agnostic about what underlies variation in measured output, which could be due to heterogeneity in fixed teacher productivity types and/or unobserved actions.

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

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

where α and $(1-\alpha)$ are her weights on not making Type I and II errors, respectively.[12] 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

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

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

[12] 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.

a Type II error (i.e., $\alpha < 1/2$). Alternatively, a high value of α may be more appropriate for an administrator allocating performance bonuses from a tight budget. If $\alpha = 1 - \alpha = 1/2$ the administrator values Type I and II errors equally.

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

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

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

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

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, c^{*EB} solves

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

The optimal cutoff for the fixed effects estimator c^{*FE} solves (3), 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, i.e., $v_{CP}^{FE}(\kappa) = v_{CP}^{EB}(\kappa) = 1$ for all desired cutoffs κ (Appendix B.2).

Remark 2 (Full information)**.** *This analysis assumes the administrator chooses a cutoff policy based on only test score information, e.g., she cannot directly condition on class size. The simplicity of such a policy makes it of obvious policy relevance, as is shown in Appendix Table 3, which documents existing incentive pay programs and shows that none condition on class size (which would be required to fix the misspecification). Additionally, when compared with a policy that may also explicitly condition on class size, a test-score-based cutoff could attenuate issues of class size manipulation for the sake of affecting the administrator's posterior about the quality of a particular teacher. However, because this assumption means empirical Bayes may be misspecified, in Appendix B.1 I consider the case in which the administrator can also directly condition on class size (which must be adjusted to be non-deterministic functions of teacher quality for the problem to remain nontrivial). Intuitively, the administrator would do no worse with this extra information, as she could always choose to ignore it. Because the obvious answer obtained in this scenario renders it of limited theoretical interest, the analysis in this paper focuses on estimators and policies that do not directly condition on class size.*

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 relative value of the 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, which are respectively α and $1 - \alpha$.

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., $v_{CP}^{FE}(\kappa) = v_{CP}^{EB}(\kappa)$). □

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

Remark 3. *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.

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

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

Proof. Because λ is monotonic (indeed, strictly increasing) in n, to simplify the proof's exposition I parameterize the empirical Bayes weights $\lambda(\cdot)$ directly as a function of θ, by assuming there is one slope for the relationship between teacher quality and weight below the population mean (β_-) and another slope for the relationship above the population mean (β_+), where either slope can be positive, negative, or zero; the result is not sensitive to the linearity assumed here. I also set $\sigma_{\bar{\epsilon}} = 1$ for all teachers for the proof of the current proposition, which does not drive the result;[13] regardless, $\sigma_{\bar{\epsilon}}$ varies between teachers in the quantitative results. The empirical Bayes weight is then

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

Suppose $\kappa < 0$ and that $c^{*EB} < 0$. Differentiating the administrator's value with respect to β_-, we obtain

$$\frac{\partial v_{CP}^{EB}}{\partial \beta_-} = (1-\alpha) \left[\int_{-\infty}^{\kappa} \frac{-c^{*EB}\theta}{(\delta_- + \beta_- \theta)^2} \phi \left(\frac{c^{*EB}}{\delta_- + \beta_- \theta} - \theta \right) dF(\theta | \theta < \kappa) \right]$$
$$+ \alpha \left[\int_{\kappa}^{0} \frac{c^{*EB}\theta}{(\delta_- + \beta_- \theta)^2} \phi \left(\frac{c^{*EB}}{\delta_- + \beta_- \theta} - \theta \right) dF(\theta | \kappa \leq \theta < 0) \right],$$

because $\frac{\partial v_{CP}^{EB}}{\partial c^{*EB}} \times \frac{\partial c^{*EB}}{\partial \beta_-} = 0$ due to the Envelope Theorem. The first term is negative because $-c^{*EB}\theta < 0$ for $\theta < \kappa$. Analogously, the second term is positive. Each term is the conditional mean of $\frac{c^{*EB}\theta}{(\delta_- + \beta_- \theta)^2}$, weighted by the density $\phi \left(\frac{c^{*EB}}{\delta_- + \beta_- \theta} - \theta \right)$. If α is not too extreme, the

[13] Details are available upon request.

first term typically 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 is the increase in class size from teacher quality. Analogously, if $\kappa > 0$ and $c^{*EB} > 0$ then by differentiating the administrator's value with respect to β_+ we can see that the administrator's value is increasing in β_+, meaning that increasing the weight associated with teacher fixed effects for teachers above the population mean improves the administrator's value.

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, suppose we started from a constant class size, i.e., $\lambda(\theta) = \delta$; if we then shifted $\beta_- > 0$ and $\beta_+ < 0$, the fixed effects estimator would provide the administrator with higher expected utility. □

Note that interior convex combinations of the estimators will not be optimal; this can be seen by manipulating σ_ϵ, which does not affect the preferred estimator. Intuitively, if one estimator is better at identifying certain teachers than the other, an intermediate estimator would also be outperformed by the corner.

Figure 1 illustrates Proposition 2 by plotting the expected utility of the administrator under the fixed effects estimator (solid red line) and the empirical Bayes estimator (dotted blue line) as a function of the cutoff policy for each estimator (x-axis), where class size is increasing in teacher quality, i.e., $\beta_-, \beta_+ > 0$, meaning that lower-quality teachers are weighted closer to the population mean than higher-quality teachers. Each curve traces out the administrator's expected utility as a function of cutoff policies, given an exogenous desired cutoff quality κ. The utility-maximizing cutoff policy for each estimator is indicated by a vertical line $c^{*\text{estimator}}(\kappa)$, where the administrator's value from using that estimator, $v_{CP}^{*\text{estimator}}(\kappa)$, is the maximum of each curve. If the administrator desires to separate the lowest quality teachers from the rest (Figure 1a), the re-weighting inherent in the empirical Bayes estimator can actually reverse teacher rankings and lead to a lower expected objective for the administrator than when the fixed effects estimator is used. The opposite is true for when the administrator wishes to separate the top teachers from the rest (Figure 1c); here, the peak of the empirical Bayes curve is higher. Intuitively, the empirical Bayes estimator is now dilating the estimated teacher quality further than the fixed effects estimator, reducing the probability the administrator makes a ranking error. When the administrator only desires to separate the upper and lower half quality teachers (Figure 1b), fixed effects and empirical Bayes both obtain the same maximum height, i.e., they return the same expected objective. An increase in either δ_- or δ_+ corresponds to an increase in the signal-to-noise ratio. Intuitively, an increase in the signal provided by student test scores increases λ,

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

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

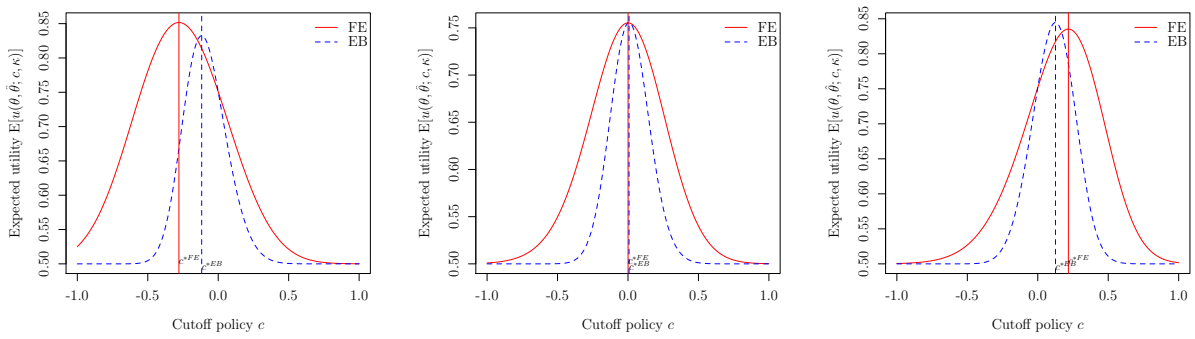
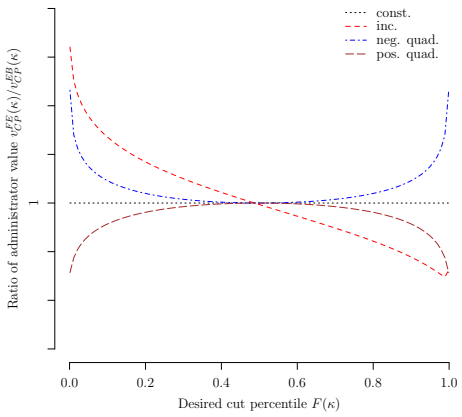


Figure 2: Difference between administrator's objective under fixed effects and empirical Bayes, by class size scenario and desired cut point



reducing the dependence of the weight on teacher quality.

Figure 2 summarizes the theoretical results for the cutoff model by comparing the performance of the estimators by plotting the ratio in value functions for the administrator ($v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa)$) as a function of the desired cut percentile $F(\kappa)$ (x-axis), for scenarios where class size is constant, increasing in teacher quality, negative quadratic in teacher quality, and positive quadratic in teacher quality (average class size is the same across scenarios). For simplicity, α has been set to 1/2; Appendix B.3 shows this does not drive the findings for the vast majority of α values. Of course, if α took an extremely high value (i.e., $\alpha \to 1$), the second term in $\frac{\partial v}{\partial \beta_-}$ above would dominate; intuitively, if the administrator did not value correctly identifying teachers below κ, their value would increase in β_-.

For each κ, estimator, and class size scenario, I solve for the administrator's optimal

cutoff policy and plug it into her objective, returning $v^{\text{estimator}}(\kappa)$. The vertical axis then plots $v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa)$ corresponding to the desired cutoff associated with the desired cut percentile $F(\kappa)$. As shown before, when class size is constant (dotted black line), the empirical Bayes cutoff is just a scaled version of the fixed effects cutoff and the administrator's value is the same under fixed effects and empirical Bayes estimators—i.e., $v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa) = 1$ for all κ. When class size is increasing in teacher quality (short-dashed red 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 1a), while the empirical Bayes estimator performs better when the administrator wishes to isolate high-quality teachers (Figure 1c). When class size has a negative-quadratic relationship with teacher quality (dot-dashed blue 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 2, the fixed effects estimator outperforms the empirical Bayes estimator at both the lowest and highest desired cutoffs. Finally, when class size is a positive-quadratic function of teacher quality (long-dashed brown line), the opposite is true. Figure 2 also demonstrates that the difference between the performance of fixed effects and empirical Bayes estimators decreases the closer the desired cut point is to the population mean of 0. Intuitively, there is less of a difference between both the estimates resulting from the fixed effects and empirical Bayes estimators when the administrator seeks to identify teachers as being on either side of the population mean (see Proposition 8 in Appendix B.4 for a proof that the administrator would be indifferent if her problem is *symmetric*).

Remark 4 (Pay for percentile)**.** *The cutoff model could be applied to a tournament-based scheme e.g., "pay-for-percentile" (Barlevy and Neal (2012)) by considering an arbitrarily large sequence of desired cutoffs and associated bonuses for being above them, which means the above results are also relevant for practitioners considering the design of such schemes or other, potentially continuous and nonlinear, ones.*[14]

Remark 5 (One-period model)**.** *The model developed here is for one period. Although using an arbitrarily large number of periods when attempting to classify teachers would increase estimator precision and, hence, administrator value, doing so would preclude using schemes for many important decisions, such as termination of extremely low-quality inexperienced teachers.*

[14]Further note that one would also want to take into account the possibility of "differential shrinkage" when finding a comparable set of teachers, which is required by the scheme derived by Barlevy and Neal (2012).

4 Asymmetric Information Models

4.1 Hidden Type Model

This section develops a hidden type, or adverse selection, model.[15] 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 may condition on quality signals but not directly on other data, e.g., class sizes. Let I denote the set of potential teachers, or applicants, who are indexed by i. Per-student output from slot j being filled by teacher i in period t is $q_{it} = \beta_0 + \theta_{i(j,t)}$, where θ_i is teacher i's quality and output for slot j is zero if it has not been assigned a teacher (i.e., $i(j,t) = \emptyset$). The quality of applicants for teaching positions is distributed according to $\theta_i \sim \mathrm{N}(\mu, \sigma_\theta^2)$, where, as in the cutoff model, $\mu = 0$. Any teacher i in the applicant pool would accept a teaching job if offered a wage at least as high as $\underline{w}$. As in Staiger and Rockoff (2010), there is an arbitrarily large number of teachers for each slot, which is not very restrictive because changes 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 $|H_{it}|$.

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

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:

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

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

Theoretical Results For simplicity, assume $\beta_0 = 0$ and set $\underline{w} = 0$.[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 (4) 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 (4) 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

[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.

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. □

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 (4) is quite general, which complicates obtaining theoretical results about how the administrator would prefer to measure teacher quality and relating results from the hidden type model to those from the cutoff-based model. Therefore, in Section 4.1.2 I study Model HT-0, a version of Model HT-G with two periods and constant class sizes. Model HT-1, in 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 $\bar{\epsilon}_{jit} \sim \mathrm{N}(0, \sigma_\epsilon^2/n)$ and $\mathrm{E}[\bar{\epsilon}_{jit}|q_{jit}] = \mathrm{E}[\bar{\epsilon}_{jit}] = 0$. 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

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

Taking expectations over the signal $\hat{q}$, we can write the administrator's value of using esti-

mator $\hat{q}$ with replacement cost χ as

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

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

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

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

We can characterize the marginal signal $\underline{q}^{*FE}$ by noting the administrator would be indifferent between replacing or retaining a teacher with that signal. The administrator's expected utility from replacing slot j's teacher is $\mathrm{E}[\theta] - \chi = -\chi$ 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 $-\frac{\chi}{\rho} \equiv \underline{q}^{*FE} > \hat{\theta}_{i(j,1)}^{FE}$. To see this, first suppose that $\chi = 0$, in which case the marginal teacher is of average quality of the existing stock of teachers; since hiring in the first period is random from the pool of applicants this means any teacher with quality expected to be below the population average (μ) would be replaced; increasing χ would lower this threshold.

4.1.3 Relation Between Preferred Estimator in Cutoff and Hidden Type Models

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. Given the results from the cutoff model, when characterizing the optimal estimator I consider the corners of the set of estimators contained by the fixed effects and empirical Bayes estimators.

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

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

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

$$\begin{aligned} v^{EB}_{HT0}(\chi) &= \max_{\underline{q}^{EB}} \Phi\left(\frac{\underline{q}^{EB}}{\sigma_{\hat{\theta}^{EB}}}\right)(-\chi) + \left(1 - \Phi\left(\frac{\underline{q}^{EB}}{\sigma_{\hat{\theta}^{EB}}}\right)\right)\rho\frac{\sigma^2_\theta}{\sigma_{\hat{\theta}^{EB}}}\frac{\phi(-\underline{q}^{EB}/\sigma_{\hat{\theta}^{EB}})}{\Phi(-\underline{q}^{EB}/\sigma_{\hat{\theta}^{EB}})} \\ &= \max_{\underline{q}^{EB}} \Phi\left(\frac{\underline{q}^{EB}}{\rho\sigma_{\hat{\theta}^{FE}}}\right)(-\chi) + \left(1 - \Phi\left(\frac{\underline{q}^{EB}}{\rho\sigma_{\hat{\theta}^{FE}}}\right)\right)\rho\frac{\sigma^2_\theta}{\rho\sigma_{\hat{\theta}^{FE}}}\frac{\phi(-\underline{q}^{EB}/(\rho\sigma_{\hat{\theta}^{FE}}))}{\Phi(-\underline{q}^{EB}/(\rho\sigma_{\hat{\theta}^{FE}}))} \end{aligned} \tag{8}$$

where the second line follows because $\sigma_{\hat{\theta}^{EB}} = \rho\sigma_{\hat{\theta}^{FE}}$. Then, if $\underline{q}^{*FE}$ solves (7) then $\underline{q}^{*EB} = \rho\underline{q}^{*FE}$ must solve (8) and, notably, return the same value for the administrator, i.e., $v^{FE}_{HT0}(\chi) = v^{EB}_{HT0}(\chi)$. □

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. For brevity, this model is developed and analyzed in Appendix C.2.

The results from Model HT-2 are strikingly similar to those from the cutoff model: (i) the administrator's preferred estimator depends on $n(\theta)$ (same as in the cutoff model), (ii) the preferred estimator does not depend on the specific parameterization of HT-2, 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. In sum, 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.

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

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

4.2 Hidden Action Model

Many teacher incentive schemes are predicated on inducing higher effort levels from teachers. This section therefore presents the workhorse CARA-Normal model of moral hazard, as developed in Bolton and Dewatripont (2005), to illustrate the potential role choice of estimator may play in affecting output in a hidden action setting. The solution of this model is the same as that in Hölmstrom and Milgrom (1987), which shows that the optimal contract features an end-of-period payment linear in measured output. This section uses the exposition of Mehta (2017), which calibrates this model to quantify the potential gains resulting from implementation of the optimal contract, using fixed effects. For convenience, I begin by sketching the model here.

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

Hölmstrom and Milgrom (1987) show that it is optimal for the administrator to pay the teacher based on the noisy output measure using a linear contract $w = \beta_0 + \beta_1 \hat{q}$, where β_1 is the share of measured output paid to the teacher. Note that, from the teacher's perspective, uncertainty comes from the composite error $\nu_i + \overline{\epsilon}_i$, which are collected as η_i. We can then write the wage as $w(a, \eta)$, where the administrator can only observe $a + \eta$. Ex-ante, teachers face the same uncertainty about η_i.[17]

Substituting for output and output measure and using the result that the optimal contract will be linear in observed output, the administrator's problem is

$$\begin{aligned} &\max_{\beta_0,\beta_1} \ \mathrm{E}_{\nu,\eta}\left[a + \nu - w(a,\eta)\right] && (9)\\ &\text{s.t.} \quad w(a,\eta) = \beta_0 + \beta_1(a+\eta) \\ &\qquad \mathrm{E}_\eta\left[-e^{-\xi(w(a,\eta)-\psi(a))}\right] \geq \underline{u} && \text{(IR)} \\ &\qquad a \in \arg\max \ \mathrm{E}_\eta\left[-e^{-\xi(w(a,\eta)-\psi(a))}\right], && \text{(IC)} \end{aligned}$$

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)

[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.

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

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 (9) when using output measure $\hat{q}_i^{FE}$ then it must be that $(\beta_0^{*FE}, \beta_1^{*FE}/\lambda)$ solves (9) when using output measure $\lambda\hat{q}_i$. Thus, the administrator obtains the same value from using either estimator. □

Intuitively, empirical Bayes contains the same amount of information as fixed effects when class sizes are constant, meaning the contract slope would simply adjust to take into account its shrunken distribution. 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.

Model HA highlights the bias-variance "tradeoff" that has potentially been the source of confusion, leading to the adoption of shrinkage estimators in many applications. If the variance of the fixed effects estimator increased, the resulting optimal contract would partially protect a risk-averse teacher by making incentives weaker in the output measure (i.e., test scores), or reducing the slope of the linear contract β_1. The more risk-averse the teacher, the more protected they would be (i.e., the shallower the slope β_1). Crucially, the optimal contract would not respond to an increase in noise by "changing the data" (e.g., switching to a lower-variance estimator), but rather, would in equilibrium adjust the way in which the data were used in remuneration (i.e., decrease β_1).

[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.

The fact that Proposition 6 shows we can re-scale the empirical Bayes estimator when class size is constant suggests the use of a biased, yet lower-variance estimator could be modeled by increasing the effective error variance σ_η^2. We can apply the informativeness principal of Hölmstrom (1979), which relates the value of a signal to how much information it contains, and rank estimators based on how much information they contain about teacher quality. I do this by examining the signal-to-noise ratio in the output measure.

Proposition 7. *The administrator's preferred estimator in Model HA depends on the relationship between teacher quality and class size.*

Proof. The empirical Bayes signal is $\hat{q}_i^{EB} = \lambda_i \hat{q}_i = \lambda_i \theta_i + \lambda_i \bar{\epsilon}_i$, which has a mean amount of signal about θ (i.e., fraction of variation explained by θ) of

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

where the numerator is smaller when $n(\theta)$ is negative quadratic and larger when $n(\theta)$ is positive quadratic. □

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.

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 discusses calibration of the additional parameters of the hidden action model and computes how choice of estimator would affect output there. Although these incentive schemes are not currently in place in Los Angeles, these exercises can serve as a useful benchmark for how the estimators might

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

perform when used in similar incentive schemes. Indeed, the fact that a high-stakes scheme was not in place obviates addressing the potential strategic re-assignment of students to teachers.

5.1 Calibration

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

Variances Schochet and Chiang (2012) compile estimates of the variances from a large number of studies in their study of error rates in value-added models, providing a good source for typical values for σ_θ^2 and σ_ϵ^2 (see 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

[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.

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 3 plots non-parametric regressions (solid blue lines) of class size on estimated teacher value-added for Reading (3a) and Math (3b). Teachers at either end of the distribution of Reading value-added have the smallest class sizes and those in the middle of the distribution have the largest class sizes. Table 1 shows the results of regressions of teacher class size on estimated teacher quality and estimated teacher quality squared. The first two columns are for Reading and the second two are for Math. The dotted black lines on Figure 3 shows the regression line fit for models in columns (1) and (3). Columns (2) and (4) are the same as regressions in (1) and (3), respectively, but exclude teachers whose estimated quality is more than two standard deviations from the population mean, showing that the estimates from the full sample are not driven by outliers. These results indicate that class size is indeed increasing in value-added in the lowest part of the distribution and decreasing in value-added in the highest part of the distribution. The relationship is not as clear for math value-added, but the regression shows that class size first increases and then decreases for reading value-added, with a negative quadratic term for math value-added. Strikingly, 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 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.

[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.

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)

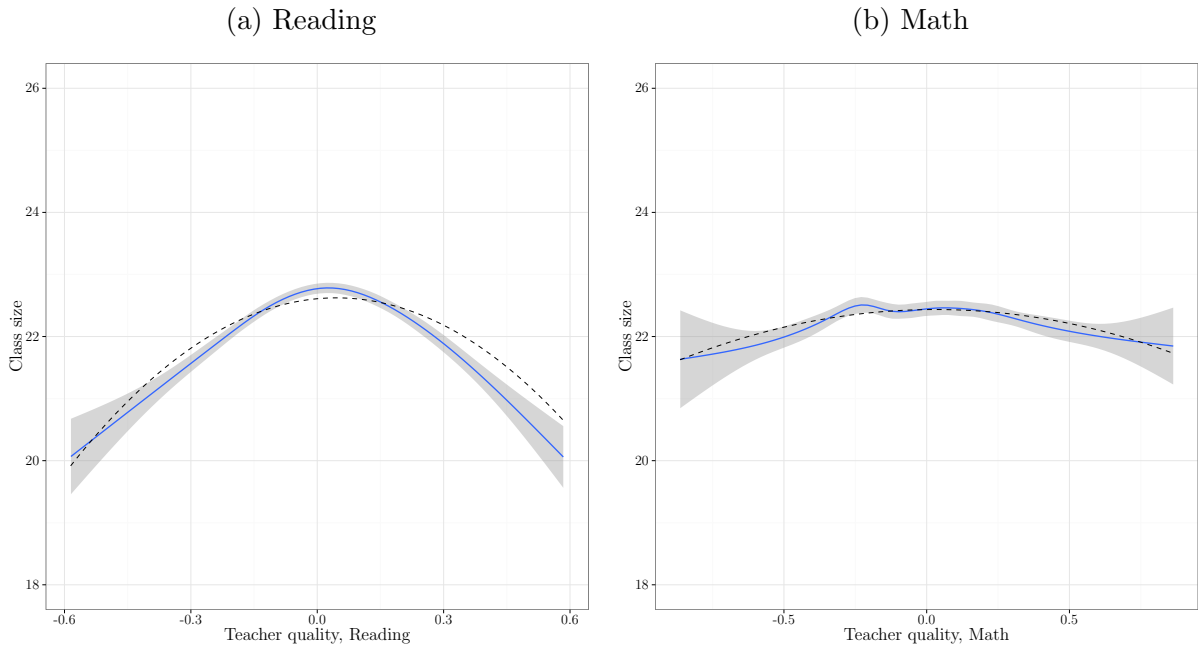
Note: ***p<0.01

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.

Figure 3: The relationship between class size and teacher quality



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 4a plots the ratio of the administrator's maximized expected objective under the fixed effects and empirical Bayes estimators ($v_{CP}^{FE}(\kappa)/v_{CP}^{EB}(\kappa)$) for Reading (solid black line) and Math (dotted red line), for desired cutoffs ranging from the lowest to the highest teacher qualities. The right panel (4b) plots how many more expected mistakes (i.e., the expected sum of Type I and II errors) the empirical Bayes estimator would make than the fixed effects estimator, assuming the Los Angeles school district employed 30,000 teachers.[25] We can see that the quadratic nature of the association between teacher quality and class size

[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`).

affects the relative performance of the fixed effects and empirical Bayes estimators in the way demonstrated by Proposition 2. The stronger negative-quadratic relationship between teacher quality and class size in the Reading test causes the larger divergence between the value of using fixed effects rather than empirical Bayes estimators. The administrator's value is higher almost everywhere when she uses the fixed effects estimator, and the relative performance of the empirical Bayes estimator is the worst at the extremes of the distribution of teacher quality. For example, the empirical Bayes estimator would make almost 800 more mistakes than fixed effects when the desired cutoff is at the 1st percentile, and 600 more when the desired cutoff is at the 99th percentile. Put another way, even when the administrator is allowed to re-optimize and choose an estimator-specific cutoff policy, using empirical Bayes would result in 9.5% more classification mistakes when the desired cutoff was at the 1st percentile of teacher quality and 7.3% more mistakes when the desired cutoff was the 99th percentile of teacher quality.[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. The performance of the fixed effects and empirical Bayes estimators most greatly diverges precisely where policies that sanction very low-performing teachers or reward very high-performing teachers would bite the most, and the fixed effects estimator returns higher expected maximized utility (i.e., in expectation would make fewer mistakes) under almost every desired cutoff.

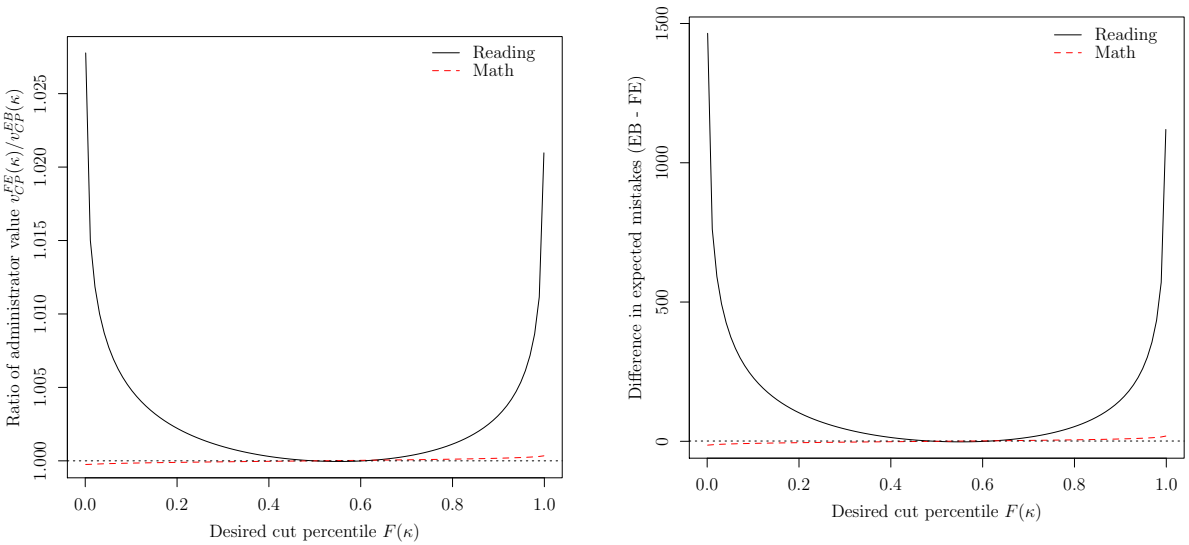
The divergence in estimator performance is largest when the desired cutoff is in the tails of true teacher quality. However, all teachers would be affected by the administrator's choice of estimator. Figure 5a plots the probability that a teacher with true quality θ, measured along the x-axis, has an estimated quality $\hat{\theta}$ above the optimal cutoff policy corresponding to a desired cutoff κ of the first percentile of true teacher quality (dotted black line), e.g., $\Pr\{\hat{\theta}^{FE} \geq c^{*FE}\}$ for the fixed effects estimator. This desired cutoff could correspond to firing teachers with quality at or below the first percentile. These probabilities are plotted for the fixed effects (solid red 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 5a. For teachers outside the shaded region, having an estimated quality below c^* would correspond to a Type I error, e.g., they were incorrectly dismissed, the probability of which corresponds to the height of the estimator-specific curve.

[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.

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

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

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



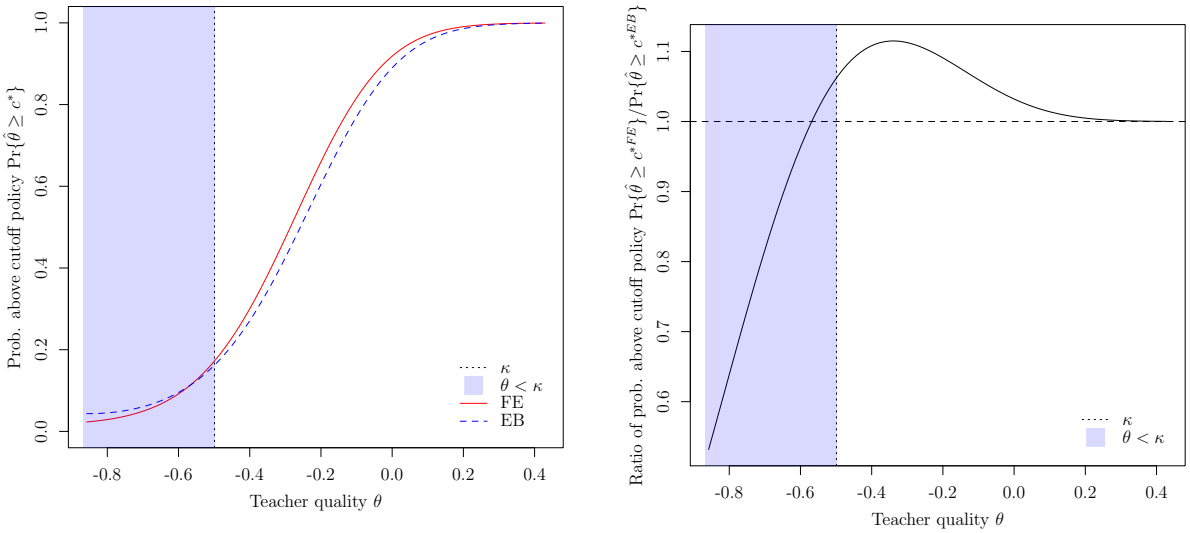
Note: Number of decisions is 30,000.

For each estimator, the probability of having estimated quality above the optimal cutoff policy increases as a teacher's true quality increases (i.e., we move to the right). However, the fixed effects estimator has a higher probability of measuring above-threshold teachers as above c^{*FE} than does empirical Bayes for its corresponding optimal cutoff policy and a lower probability of measuring below-threshold teachers as above c^{*FE}. That is, fixed effects would have lower probabilities of both Type I and Type II errors. This is more clear in Figure 5b, which plots the ratio of probability of the estimate being above the respective cutoff for fixed effects over empirical Bayes, i.e., $\Pr\{\hat{\theta}^{FE} \geq c^{*FE}\}/\Pr\{\hat{\theta}^{EB} \geq c^{*EB}\}$. 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 5: Probability of being measured above optimal cutoff policy, given $F^{-1}(\kappa) = 0.01$

(a) Probability of being above c^*

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



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.

I use Model HT-2 to get a rough 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 naturally viewed in terms of teacher quality, which has been shown to appreciably affect economic output (Hanushek (2011)).

Expected output when using empirical Bayes and the optimal reservation signal policy $\underline{q}^{*EB}(\chi, n(\theta))$, is 0.058; teachers with quality measures in the bottom 36% would be replaced That is, second-period teacher quality from using empirical Bayes would be 5.8% of a standard deviation larger 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% larger. If instead, we used the value $\chi = 0.07$ from Staiger and Rockoff (2010), the reservation signal would be lower, in response to the larger replacement cost; here, teachers with the lowest 32% signals would be replaced. Expected teacher quality in the second period would be 5.6% of a standard deviation larger when using empirical Bayes than it would be in a world where all teachers were retained, and 0.22% larger under fixed effects than when using empirical Bayes.

5.4 Quantitative Findings: 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.

Therefore, this section takes two approaches to roughly examine how choice of estimator might affect optimal output in a hidden action environment. First, it uses estimates from Muralidharan and Sundararaman (2011) to calibrate parameters from the hidden action model. Second it computes the effect on output from using either estimator of teacher quality for a wide range of model primitives. The approaches use the relationship between class size and teacher quality for Reading, from Section 5.1[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 absent output-based incentives.

In the hidden action model, output is a function of the action, which itself depends on the variance of noise η, CARA parameter ξ, and cost parameter γ. I first characterize how much information the administrator can extract about teacher quality (here, teacher effort choices) using either estimator. I do this by calibrating the implied variance of the composite error η for the fixed effects and empirical Bayes estimators (details are in Appendix D.3).[28] Based on Proposition 7, I model the information loss when using empirical Bayes under a negative

[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, which would understate the gain in output from using fixed effects instead of empirical Bayes.

quadratic relationship between class size and teacher quality by increasing the measurement error variance on teacher action, σ_η^2, by 3.2%.[29]

Mehta (2017) uses data from Muralidharan and Sundararaman (2011), which estimates the effect of a linear output-based incentive scheme for teachers in the Indian state of Andhra Pradesh, and other information to calibrate the model parameters $(\gamma, \xi, \sigma_\eta^2)$. These are then used to characterize the optimal contract when using the fixed effects estimator and the gains from implementing the optimal contract, which at the calibrated parameters would be over six times larger than those in Muralidharan and Sundararaman (2011). Here, I take that calibration as given and compute the effect of using the empirical Bayes estimator on equilibrium output under the optimal contract.

Briefly, Mehta (2017) exploits the teacher's optimal choice of action, which solves (IC) in (9) but does not rely on optimality of the slope β_1, to map (β_1, a) to the cost $\gamma = 4.385 \times 10^{-5}$. The CARA parameter is set to $\xi = 6.7 \times 10^{-3}$, the mean estimated CARA from the benchmark model of Cohen and Einav (2007), Table 5. Finally, the variance of output is calibrated to σ_η^2 =6,076,631\$^{2}. At the calibrated parameters, the optimal slope is $\beta_1^{*FE} = 0.483$, which has a corresponding optimal action of $a^{*FE} = \$11,011.34$; this corresponds to an average increase in student achievement of 0.919 sd. Either increase is more than six times larger than the estimated increase in student achievement stemming from the much weaker incentives provided under the experiment.

In contrast, using the above reckoning that empirical Bayes increases the variance of η by 3.2%, using empirical Bayes would produce an optimal slope of $\beta_1^{*EB} = 0.475$ 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 would affect output for all teachers.

Sensitivity Analysis Via Parameter Grid The mean class size in the Los Angeles data is 22.5, much smaller than the mean of 37.5 used in the above calibration. Smaller class sizes would increase the variance of the output measure. Moreover, Dohmen and Falk (2010) document that teachers are more risk-averse than other workers. Therefore,

[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.

it would seem reasonable to examine how estimator-specific output would be affected by varying the parameters of the hidden action model. Mehta (2017) does this for a grid of points covering a wide range of alternative values of σ_η^2 and ξ, ranging from one half to ten times the calibrated value of each parameter.[30] Briefly, as teachers become more risk averse (increasing ξ) or the output measure becomes noisier (increasing σ_η^2), both optimal incentive strength and output would decrease. For example, the increase in output ranges from over 1.5 sd in student achievement to around 0.5 sd when teachers are ten times more risk averse than their calibrated value of $\xi = 6.7e-3$; this latter figure is only about three times the estimated effect of the incentive scheme.

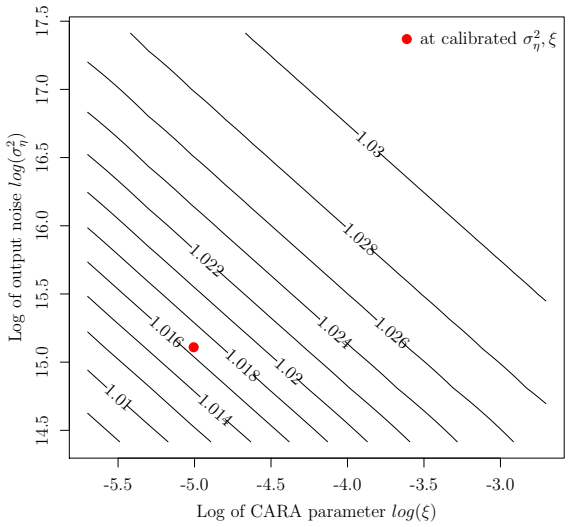
As interesting as these results may be in their own right, the goal here is to quantify the difference in output stemming from using one estimator versus another. Figure 6 presents a contour maps of the ratio in optimal output from using fixed effects over that using empirical Bayes. Although, as just discussed, optimal incentive strength and output gains vary considerably with respect to σ_η^2 and ξ, the output gain associated with using fixed effects versus empirical Bayes ranges from just above 1% to around 3%. Intuitively, the higher noise in empirical Bayes matters more (relative to the cost γ) when teachers are more risk averse or when the baseline variance on the shock to output is higher. Of course, we cannot know the exact amount by which the output would be lower were the administrator to use empirical Bayes; knowing this would require the development and estimation of a richer structural model, a promising avenue for future research. However, the variable share of compensation, calculated in Mehta (2017), can provide further guidance. As with the slope and output, this share declines as the output noise variance and degree of risk aversion increase. Suppose it seemed reasonable that, in the optimal arrangement, the variable share of compensation for teachers would be at most around 2% of their income. Then the gain in output from switching from empirical Bayes to fixed effects would be about 2-3%, which is even larger than it was at the calibrated parameter values.

6 Discussion

While economic theory can help inform education policy, measurement issues are also important when considering how to actually use data. Due to their statistical properties, empirical Bayes estimators of teacher value-added are used by many education researchers and prac-

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

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



titioners to make inferences about teacher quality, which may serve as inputs to high-stakes decisions like bonus assignments, personnel decisions, or wages. More generally, shrinkage estimators are used on a broad array of policy-relevant applications. It is not obvious this should be the case. If an estimator is going to be used to make a decision, then studies of its bias and other statistical properties are certainly useful, but as an intermediate—not final—step in their evaluation.

In this paper, I show that the preferred estimator depends on information that is plausibly part of an administrator's context. The preferred estimator would be the same for wide ranges of underlying parameters for all the models considered and is determined by the relationship between class size and teacher quality. Because it is possible to compute the preferred estimator without knowing the specific parameterization of the relevant model, 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 those with calibrated parameters (e.g., Rothstein (2014)).

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. Suppose an administrator had been using empirical Bayes in an incentive scheme. Would it make sense to switch to fixed effects? The relevant comparison, from an economic perspective, is a cost-benefit one. It is important to note that the intervention considered in this paper is very

easy to implement and virtually costless—to use a different, more transparent estimator of teacher quality—and that the preferred estimator would be the same across several models of the administrator's objective. Indeed, in all likelihood, the relative cost of using fixed effects would be zero, or even negative, given the increased transparency of fixed effects, which could translate to a lower nonpecuniary cost incurred by society. Then, by an economic criterion, these results suggest an obvious benefit from using fixed effects instead of empirical Bayes in the design of teacher incentive schemes if, as was suggested previously, class size is negative quadratic in teacher quality in the relevant context. Administrators hesitant to implement incentive schemes may take comfort in at least knowing schemes were better-designed for their purposes, reducing the change of public backlash. It is important to note that the results are not just driven by the econometric endogeneity misspecification (i.e., class sizes that are negative quadratic in teacher quality). Absent this econometric endogeneity, empirical Bayes would still not return a higher value to the administrator, as uniform shrinkage would be undone by the administrator's re-optimization.

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.

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

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

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 schemes 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

$$u_{CP}(\theta, \hat{\theta}; c, \kappa) = \alpha 1\{\hat{\theta} \geq c \cap \theta \geq \kappa\} + (1-\alpha)1\{\hat{\theta} < c \cap \theta < \kappa\} \quad \xrightarrow{p} \alpha 1\{\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)1\{\theta < c \cap \theta < \kappa\}, \tag{10}$$

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

$$\begin{aligned} plim_{\sigma_{\bar{\epsilon}} \to 0} u_{CP}(\theta, \hat{\theta}; c, \kappa) &= \alpha 1\{\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)1\{\theta < c \cap \theta < \kappa\} \\ &= \alpha 1\{\lambda(\theta)\theta \geq c \cap \theta \geq \kappa\} + (1-\alpha)1\{\lambda(\theta)\theta < c \cap \theta < \kappa\}, \end{aligned} \tag{11}$$

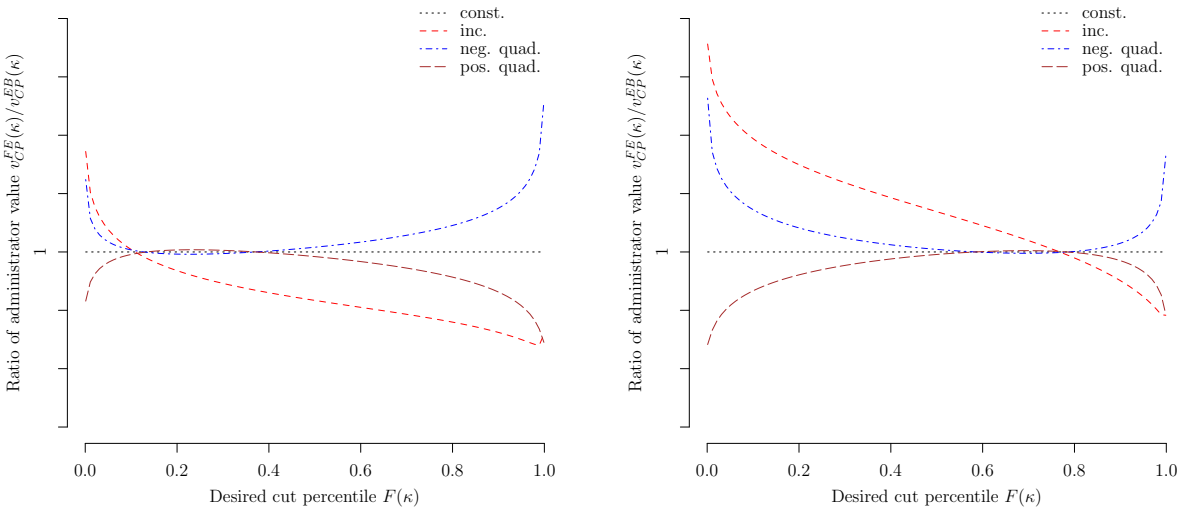
which is maximized at $c = \kappa/\lambda(F^{-1}(\kappa))$. The probabilities of the events in both (10) and (11) are all 1, giving an expected utility of 1 for all teacher qualities, which then integrates to a value of 1 for each estimator.

B.3 Asymmetric Type I and Type II Weights

The administrator's preferred estimator is not sensitive to the assumption that α is close to 1/2. Figure 7 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 7a 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 7b 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 7: 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 8

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 8. *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, and 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. Also, 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 (7) must be adapted to obtain the administrator's value from using fixed effects:

$$v_{HT2}^{FE}(\chi) = \max_{\underline{q}^{FE}} \left(\int_{-\infty}^{\infty} \Phi\left(\frac{\underline{q}^{FE} - \theta}{\sigma_{\bar{\epsilon}}(n(\theta))} \right) dF(\theta) \right) (-\chi) + \int_{-\infty}^{\infty} \left(1 - \Phi\left(\frac{\underline{q}^{FE} - \theta}{\sigma_{\bar{\epsilon}}(n(\theta))} \right) \right) \left(\frac{\sigma_\theta^2}{\sigma_{\hat{\theta}^{FE}(n(\theta))}} \frac{\phi\left(-\underline{q}^{FE} / \sigma_{\hat{\theta}^{FE}(n(\theta))} \right)}{\Phi\left(-\underline{q}^{FE} / \sigma_{\hat{\theta}^{FE}(n(\theta))} \right)} \right) dF(\theta), \tag{12}$$

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 9. The administrator's value from using the empirical Bayes estimator is

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

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

Proposition 3 would still apply here. The estimator-specific reservation signals, $\underline{q}^{*FE}$ and $\underline{q}^{*EB}$, are respectively obtained by numerically solving (12) and (13).

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.

Proposition 9. *In Model HT2, the administrator's preferred estimator depends on the relationship between teacher quality and class size.*

Proof. As in the cutoff model, set $\sigma_{\bar{\epsilon}} = 1$, which does not drive the result,[31] and parameterize the empirical Bayes weights via

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

also set $\sigma_{\hat{\theta}^{FE}(n(\theta))} = \sigma_\theta = 1$ to simplify exposition, resulting in the administrator's value

$$\begin{aligned} v^{EB}_{HT2}(\chi) = &\left(\int_{-\infty}^{\infty} \Phi\left(\underline{q}^{*EB}/\lambda(n(\theta)) - \theta\right) dF(\theta)\right)(-\chi) \\ &+ \int_{-\infty}^{\infty} \left(1 - \Phi\left(\underline{q}^{*EB}/\lambda(n(\theta)) - \theta\right)\right) \frac{\phi\left(-\underline{q}^{*EB}/\left(\lambda(n(\theta))\right)\right)}{\Phi\left(-\underline{q}^{*EB}/\left(\lambda(n(\theta))\right)\right)} dF(\theta). \end{aligned} \tag{14}$$

Note that $\underline{q}^{*EB} < 0$ if $\chi > 0$, leading us to examine the role played by β_-. Differentiating with respect to β_-, we obtain

$$\begin{aligned} \frac{\partial v^{EB}_{HT2}(\chi)}{\partial \beta_-} = &-\chi \int_{-\infty}^{0} \frac{\underline{q}^{*EB}\theta}{(\delta_- + \beta_-\theta)^2} \phi\left(\underline{q}^{*EB}/(\delta_- + \beta_-\theta) - \theta\right) dF(\theta|\theta < 0) \\ &- \int_{-\infty}^{0} \frac{\underline{q}^{*EB}\theta}{(\delta_- + \beta_-\theta)^2} \phi\left(\underline{q}^{*EB}/(\delta_- + \beta_-\theta) - \theta\right) \frac{\phi\left(-\underline{q}^{*EB}/\left(\delta_- + \beta_-\theta\right)\right)}{\Phi\left(-\underline{q}^{*EB}/\left(\delta_- + \beta_-\theta\right)\right)} dF(\theta|\theta < 0) \\ &+ \int_{-\infty}^{0} \left(1 - \Phi\left(\underline{q}^{*EB}/(\delta_- + \beta_-\theta) - \theta\right)\right) \frac{\partial\left[\frac{\phi\left(-\underline{q}^{*EB}/(\delta_- + \beta_-\theta)\right)}{\Phi\left(-\underline{q}^{*EB}/(\delta_- + \beta_-\theta)\right)}\right]}{\partial \beta_-} dF(\theta|\theta < 0) \end{aligned} \tag{15}$$

because $\frac{\partial v^{EB}_{HT2}}{\partial \underline{q}^{*EB}} \times \frac{\partial \underline{q}^{*EB}}{\partial \beta_-} = 0$ due to the Envelope Theorem. The first line is negative because $\chi > 0$ and $\underline{q}^{*EB}\theta > 0$. The second line includes the Inverse Mills Ratio term $\frac{\phi\left(-\underline{q}^{*EB}/(\delta_- + \beta_-\theta)\right)}{\Phi\left(-\underline{q}^{*EB}/(\delta_- + \beta_-\theta)\right)}$, i.e., the expected value of θ given the signal was above the reservation value. Therefore, the second line is also negative. The third line contains $\frac{\partial\left[\frac{\phi\left(-\underline{q}^{*EB}/(\delta_- + \beta_-\theta)\right)}{\Phi\left(-\underline{q}^{*EB}/(\delta_- + \beta_-\theta)\right)}\right]}{\partial \beta_-}$. To evaluate the sign of this expression, note that $\lim_{\chi \to 0} \underline{q}^{*EB} = 0$ and consider a very small value of χ (allowing us to abstract from changes in $\underline{q}^{*EB}$) and a large, positive, β_-, which would shrink measured quality toward zero by a larger amount, the lower true quality was.

[31] Details are available upon request.

Since we are considering β_-, the differential shrinkage would exclusively be coming from true qualities below the reservation value; intuitively, having a signal above $\underline{q}^{*EB}$ is not as good news about underlying quality when $\beta_- > 0$. If we instead decreased β_-, then higher signals would be better news about the corresponding expected output; therefore, this derivative will also be negative. Putting this together, the derivative of administrator value with respect to β_- is negative, meaning that a negative quadratic $n(\theta)$ would result in the administrator preferring fixed effects and a positive quadratic relationship would have the administrator prefer empirical Bayes. □

Note that the reservation policy is decreasing in χ. For very large replacement costs, which correspond to low $\underline{q}^{*EB}$, the third line may be positive since differential shrinkage toward the population mean would now comprise better news about underlying quality. Therefore, I numerically solve for the value according to each estimator for a wide range of replacement costs and under different class size scenarios: constant, negative quadratic, and positive quadratic.[32] Figure 8 shows how the relationship between class size and teacher quality would affect outcomes in HT-2. The left panel, Figure 8a, 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 8b, 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

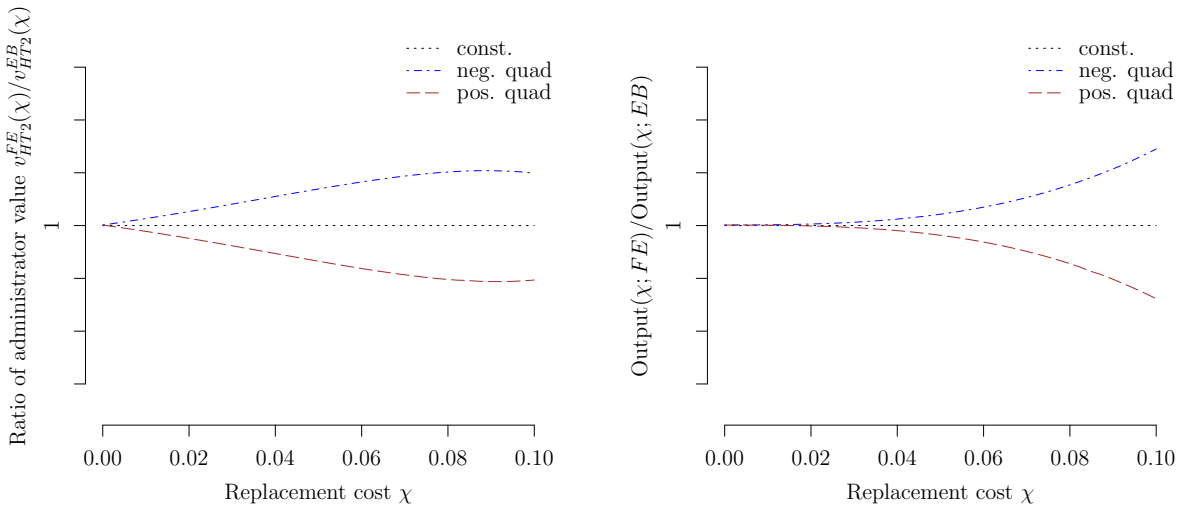
[32]Wiswall (2013) reports that teachers with 30 years of experience have value-added that is one standard deviation higher than new teachers and 0.75 standard deviations higher than teachers with five years of experience; this implies a 0.25 sd difference acquired in the first five years of experience. 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.

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 8 then likely encompass costs for multi-period environments as well.

Figure 8: 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



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 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_θ^2 and σ_ϵ^2 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 $\sigma_\theta^2 = 0.046$ 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 $\sigma_\epsilon^2 = 0.953$ 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_θ^2 and σ_ϵ^2 are approximately the same if school-level variances are excluded from σ_θ^2 or class-level variances are excluded from σ_ϵ^2, 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.[33]

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

[33] 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.

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, $\overline{\epsilon}_i^{sim}$; form simulated estimated teacher quality according to $\hat{\theta}_i^{sim} = \theta_i^{sim} + \overline{\epsilon}_i^{sim}$.

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 $\left(\hat{\beta}_0^{sim}, \hat{\beta}_1^{sim}, \hat{\beta}_2^{sim}, \hat{\sigma}_e^{sim}, n_{p25}^{sim}, n_{p50}^{sim}, n_{p75}^{sim}\right)$.

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

References

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

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.

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

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.

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

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

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

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

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

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

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

Guarino, C. M., 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.

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.

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.

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

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

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

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

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

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

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

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.

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

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

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

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

Mehta, N., "The Potential Output Gains from Using Optimal Teacher Incentives: An Illustrative Calibration of a Hidden Action Model," *UWO Centre for Human Capital and Productivity Working Paper Series*, 2017(8), 2017.

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

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

Murnane, R. J., "The Impact of School Resources on the Learning of Inner City Children." 1975.

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

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

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

Rivkin, S., 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.

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

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

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

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

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.

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, "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.