Which Value-Added Estimator Works Best and When?

Elizabeth Santorella

November 14, 2017

Abstract

As value-added estimation spreads to fields outside education, where sample sizes may be small and experimental validation infeasible, estimators that perform well without millions of observations are increasingly needed. I clarify conditions under which existing methods are identified, sign their biases, and derive asymptotic standard errors, and I develop a likelihood-based estimator. In simulation, the likelihood-based estimator nearly eliminates bias without increasing variance. I compare results from each estimator on two real datasets: data on elite bureaucrat postings and local economic activity in India, and a dataset of eighth grade math teachers and test scores in New York City.

Value-added estimators have been extensively used to study teachers and other groups. These estimators describe how dispersed teachers (or others) are in their effects on an outcome: for example, variation in teacher quality contributes to about 1% of the variance in student test scores. Value-added modeling is also used by school districts to rank teachers and make firing decisions. Although a large volume of research has investigated whether and when the identification assumptions of value-added models hold (Rothstein, 2009, 2010; Koedel and Betts, 2011; Chetty *et al.*, 2014; Rothstein, 2017). the statistical properties of these estimators are less studied, especially in finite samples. For example, standard errors and hypothesis tests are often unavailable, and parameter estimates can be badly biased even when identified.

A common use of value-added modeling is to measure what portion of variance in outcomes is due to variation in teacher quality. This number is of interest because if teachers vary little in their quality, then attempts to hire and retain better teachers may have little effect on student achievement. Estimates vary: Kane and Staiger (2008), who experimentally validated their estimates, albeit with large standard errors, found that the standard deviation in teacher quality in Los Angeles was 10% of a standard deviation in test scores, while Buddin (2011) measured 27%. [1]

I think of a value-added model as one with the following properties: Each observation i corresponds to some individual $j(i)$. When forming the best linear predictor of an outcome given an indicator for $j(i)$ and covariates, the coefficient on the indicator is $\mu_{j(i)}$. These $\mu_{j(i)}$ have a causal interpretation: If a student's teacher j is experimentally replaced with teacher

[1] What constitutes a large amount of dispersion in teacher quality is contentious. If the standard deviation of teacher quality is only 10% of the standard deviation of test scores, teachers contribute only 1% of variance. On the other hand, since teachers affect many students and have persistent effects on students' income and educational attainment, the value of improving a teacher's effectiveness by one standard deviation could be quite high (Chetty *et al.*, 2014).

j', the student's outcome increases in expectation by $\mu_{j'} - \mu_j$. The μ are drawn identically and independently from the same distribution, $\mu_j \overset{iid}{\sim} F$, and the distribution F itself is of interest. The individual components of μ may also be of interest. High-dimensional covariates and few observations for each teacher are common complications, making it inadvisable to simply estimate μ with a fixed effects regression. This setup lends itself naturally to an Empirical Bayes estimation procedure, which first estimates the distribution F and then forms a "posterior" estimate of μ. Empirical Bayes methods have been used to study teachers and in various other settings. For example, Ellison and Swanson (2016) study how much of the variation between schools in the fraction of high math achievers that are female is due to variation in schools. Feng and Jaravel (2016) study variation in patent examiners' propensity to grant patents and which patents benefit from being assigned to a lenient patent collector. Furthermore, many studies that do not rely explicitly on the teacher value-added literature share this literature's interest in estimating both individual effects and their distribution. For example, there is a wide literature in labor economics on estimating individual and firm effects (i.e. Abowd *et al.* (1999)). Recently, Barnett *et al.* (2017) studied "the extent to which individual physicians vary in opioid prescribing and the implications of that variation." Others have studied hospital effects on C-sections (Kozhimannil *et al.*, 2013) and variation in judge sentencing tendencies (Green and Winik, 2010).

In this project, I survey several popular value-added estimation procedures and study their statistical properties. I discuss conditions under which models are identified, clarify whether estimators are consistent or unbiased, and derive asymptotic, parametric standard errors. I also develop a maximum (quasi-)likelihood estimator. Monte Carlo simulations confirm theoretical predictions, and suggest that a bias-corrected maximum likelihood estimator nearly eliminates bias with no increase in variance. I compare estimates from different methods in two real real datasets: data on elite bureaucrats and local economic outcomes in India, and data on eighth grade math teachers and test scores. My focus is on the portion of variance that is due to variation in teacher quality. Although I derive formulas for individual value-added sores, I do not evaluate the accuracy of these scores. For clarity, I often use terminology relating to teachers and classrooms since value-added modeling is most used for studying teachers. However, these results extend readily to different settings.

This paper proceeds as follows. In Section 1, I develop a toy model to motivate why policymakers may care about the variance of teacher effects. In Section 2, I recap the historical development of the value-added literature and the settings in which value-added estimators have been used. Section 3 describes several estimators whose properties I develop and compare. Section 4 discusses the behavior of several procedures in Monte Carlo Data, and Section 5 compares the performance of these estimators on two real data sets: teachers' effects on test scores in New York City, and bureaucrats' effects on project completions in India. Section 6 concludes with recommendations about which estimator to use.

1 Toy Model, Motivation

Why should policymakers care about the variance of teacher effects? This section lays out a toy model in which the variance of teacher effects is a sufficient statistic for policy decisions. In particular, there should be more investment in teacher training when teachers vary more in their quality.

Assume that student i's academic ability a_i is a linear combination of teacher quality and other inputs x. Student i has teacher j, and teacher j has quality μ_j. Teacher quality is drawn from some distribution with mean $\bar{\mu}$ and variance σ_μ^2. In other words,

$$a_i = \mu_{j(i)} + x$$
$$\mu_j \sim \left[\bar{\mu}, \sigma_\mu^2\right]$$

A high σ_μ^2 is taken as an indication that interventions that improve teacher effectiveness may be worthwhile; if, on the other hand, teachers do not vary greatly in effectiveness, improving teacher quality will be difficult. To motivate this conclusion, imagine that a policymaker can invest either in increasing x, at constant marginal cost $C(x) = x$, or in improving teacher quality, at increasing marginal cost $C(\bar{\mu}_j) = \frac{C\bar{\mu}_j^2}{2\sigma_\mu}$: marginal cost is increasing in quality, but decreasing in the variance of quality [2]. If the policymaker solves

$$\max_{\bar{\mu}} \bar{\mu} + x \quad \text{subject to} \quad C(x) + C(\bar{\mu}) \leq M,$$

they will set $\bar{\mu}^* = \sigma_\mu / k$, assuming an interior solution, and invest $\sigma_\mu / 2k$ in teacher quality. Investing in teacher quality is more valuable when teacher quality is more variable, because high variation in teacher quality is an indication that the returns to training are high.

Other models also lead to the conclusion that σ_μ^2 is important. For example, when σ_μ^2 is high, deselecting low-quality teachers and replacing them with average teachers is more beneficial, as is trying to recruit above-average teachers.

2 Literature

> The extensive investigation of the contribution of teachers to student achievement produces two generally accepted results. First, there is substantial variation in teacher quality as measured by the value added to achievement or future academic attainment or earnings. Second, variables often used to determine entry into the profession and salaries, including post-graduate schooling, experience, and licensing examination scores, appear to explain little of the variation in teacher quality so measured, with the exception of early experience (Hanushek and Rivkin, 2010).

The earliest work on teacher quality noted that teacher output appeared unrelated to observable teacher characteristics other than experience and perhaps teacher test scores, and sometimes argued that variation in teacher quality is not an important determinant of differences in educational outcomes (Hanushek and Rivkin, 2010, 2006) [3]. However, later work has focused on "outcome-based" measures of teacher quality, treating quality as a

[2] A variety of teacher quality production functions can motivate this structure. For example, suppose teachers randomly receive training $T_j \sim N(0,1)$ at constant marginal cost c. and quality is generated by $\mu_j = \sigma_\mu T_j$. Then spending $c\Delta$ to increase training by Δ increases mean teacher quality by $\sigma_\mu \Delta$.

[3] Briggs and Domingue (2011) finds that teachers' educational backgrounds do predict teacher effects

Table 1: *Estimates of the variance of teacher effects, $\hat{\sigma}^2_\mu$, and forecast bias adapted from Table 6 of Kane and Staiger (2008). "1 - forecast bias" is the coefficient from regressing experimental test scores on non-experimentally estimated value-added scores. 95% confidence intervals are in brackets.*

	Math	Math	Reading	Reading
Student FEs?	N	Y	N	Y
Var $(\mu_j)^{1/2}$	0.219	0.101	0.175	0.084
1 - forecast bias	0.905	1.859	1.089	2.144
	[0.552, 1.258]	[0.938, 2.780]	[0.523, 1.655]	[0.899, 3.389]

latent variable to be estimated, and found that teachers explain about 1% to 3% of the variance in student outcomes (Hanushek and Rivkin, 2012).

The identification requirements of value-added models that treat teacher quality as a latent variable make such models controversial. These models typically involve a sorting on observables requirement: Any association between student attributes and teacher identities must be captured by variables included in the model. This requirement is necessary both for estimating the fraction of variance in student outcomes that is due to variation in teacher quality, and for evaluating individual teachers. Sorting on observables could be violated if, for example, students assigned to better teachers have parents who push them to study hard. More subtly, imagine that all teachers are identical, but some teachers are consistently assigned high- or low-achieving students; if student achievement can't be predicted well by observables, then these teachers will appear to be the cause of their students' achievement, and teacher quality may appear to vary even when it does not. The validity of the sorting on observables requirement has been contested (Rothstein, 2010). However, in this paper I focus on issues that can arise even when identification requirements are obeyed.

Several studies have addressed whether value-added scores are "forecast unbiased": that is, whether a teacher with a value-added score of $\hat{\mu}$ causally raises test scores by $\hat{\mu}$, in expectation. Unbiased estimates of the variance of teacher effects, $\widehat{\text{Var}}(\mu)$, are necessary for forecast-unbiased value-added scores, since value-added scores are a product of a mean residual and a shrinkage factor based on $\hat{\sigma}^2_\mu$. The literature has typically interpreted forecast bias as a sign of insufficient controls for student-teacher sorting, but it can also reflect bias in $\widehat{\text{Var}}(\mu)$, an issue I consider in this paper. Randomized and quasi-experimental analyses have somewhat ameliorated concerns that sorting on unobservables biases estimates of the variance of teacher quality upwards. [4] Previous studies have generally concluded that value-added scores are close to forecast-unbiased, after converging on sets of specifications that tend to work well (Jacob, 2005; Kane and Staiger, 2008; Rothstein, 2009; Chetty *et al.*, 2014).

However, experimentally validated estimates tend to be smaller than other estimates, and methods of checking for bias are controversial. One of very few randomized assessments

[4]Kane *et al.* (2013b) find, using the Measures of Effective Teaching project, that a teacher predicted to improve test scores by 1 unit improves test scores by 0.7 units when randomly assigned to different classrooms. This discrepancy could be either because value-added scores were tainted by sorting of students to teachers, or because $\text{Var}(\mu)$ was overestimated. They estimate $\text{Var}(\mu)$ to be 2.6% to 3.2% in math and 1% to 1.4% in ELA.

The value-added methods used in the MET project are not easily comparable to other methods surveyed here, because the researchers had access to video data and teacher quality surveys.

of value-added modeling comes from Kane and Staiger (2008), who estimated estimated individual value-added scores for teachers in Los Angeles, randomly assigned students to teachers in the next year, and confirmed that the previous value-added scores were an unbiased predictor of future student achievement. The results of Kane and Staiger (2008), reproduced in Table 1, show that a teacher one standard deviation above average improves math scores by 0.219 standard deviations in models that don't include student fixed effects and by 0.101 in models that do, with analogous estimates of 0.175 and 0.084 for reading; their experimental results suggest that estimates without student fixed effects are nearly unbiased and estimates with student fixed effects significantly understate teachers' contributions to test score variation by underestimating $\widehat{\text{Var}}(\mu)$ and imposing too much shrinkage. However, they are unable to rule out large degrees of bias. Estimates of about 0.1 are relatively small for this literature. For example, Buddin (2011) also analyzed data from Los Angeles – the same district studied by Kane and Staiger (2008) – to generate value-added scores that were published in the LA Times Felch *et al.* and found that a teacher one standard deviation above average improves math test scores by 0.27 standard deviations. That is, Buddin (2011) find that teachers account for 7% of the variance in math test scores in Los Angeles, while according to Kane and Staiger (2008) they account for only 1%. Lacking experimental data, Chetty *et al.* (2014) introduce the use of "teacher switching quasi-experiments": they argue that teachers switch schools for exogenous reasons and that after switching schools, teachers' value-added will not be correlated with the ability of their current students. The quasi-experiments indicate that forecast bias is quite small: the coefficient from regressing changes in test scores with changes in value-added (with controls) is approximately 0.97 and at least 0.9. Rothstein (2017) replicates the quasi-experiments in North Carolina, questions the randomness of teacher transfers. He finds similar results when using the same specifications as Chetty, Friedman, and Rockoff, but a forecast bias of about 10% when using test score gains instead of levels as the dependent variable; he argues that this is because high value-added teachers tend to move to improving schools. On the other hand, Chetty *et al.* (2017) argue that Rothstein's specifications can generate bias, and show through simulation that it is possible to find that Rothstein's tests fail even when identified.

Despite uncertainty about how to test identification restrictions, most researchers agree that in large samples and with controls for past student test scores, value-added models can accurately estimate the variance in teacher quality. (Useful reviews are given by Koedel *et al.* (2015), Hanushek and Rivkin (2010), and Staiger and Rockoff (2010).) By contrast, using value-added models to assess individual teachers remains controversial Koedel *et al.* (2015). Briggs and Domingue (2011), for example, re-analyze data from Buddin (2011), whose results were published in the LA Times, and find that with richer controls, individual teachers' value-added scores shift dramatically. Staiger and Rockoff (2010) state that value-added scores have a reliability of 30% to 50%.

Statistically, the value-added literature has been influenced by the literature on Empirical Bayes, hierarchical linear models, and correlated random effects.

In summary, two well-studied areas are whether the identification requirements of value-added models are obeyed and how confidently these models can evaluate individual teachers. There has been relatively little work on how value-added procedures behave in finite samples and how to quantify uncertainty in structural parameters.

Table 2: *Comparison of estimators. Asymptotics are as the number of teachers approaches infinity.*

	MLE	*Bias-Corrected MLE*	*Kane and Staiger*	*Mod-KS*
Consistent under baseline model	Yes	Yes	No	Yes
Consistent under baseline + no sorting	Yes	Yes	Yes	Yes
Bias under baseline	Pos	Minimal	Neg	Yes
Closed-form solution	No	No	Yes	Yes
Closed-form asymptotic standard errors	MLE	No	GMM	GMM

3 Estimators

In this section, I lay out a statistical model and discuss estimation of that model via maximum likelihood. I then discuss two other estimators: the estimator used in Kane and Staiger (2008), and a modification to Kane and Staiger (2008)'s estimator similar to those suggested by Guarino *et al.* (2014) and Chetty *et al.* (2014), "modified-KS." Both maximum (quasi-)likelihood and modified-KS consistently estimate this model. I show that the Kane and Staiger estimator consistently estimates this model after imposing a no-sorting restriction, and is otherwise negatively biased.

Observations are at the student level. Student i has classroom $c(i)$, $j(i)$, test score y_i, and covariates $\boldsymbol{x}_i$. [5] Data is drawn from some distribution $\mathcal{D}$. (Although a likelihood function will be derived using normality assumptions, $\mathcal{D}$ need not be normal.) I describe the model in terms of best linear predictors. The model's parameters are best linear predictor coefficients and variances of teacher effects and error terms. Asymptotics are as the number of teachers approaches infinity.

To begin defining best linear predictors, stack all of the data from teacher j, who has n_s students, into a vector $\boldsymbol{y}_j \in \mathcal{R}^{n_s}$, a matrix $\boldsymbol{x}_j \in \mathcal{R}^{n_s \times k}$, and mean covariates $\bar{\bar{x}}_j \in \mathcal{R}^k$. [6] $\boldsymbol{y}_j$ and $\boldsymbol{x}_j$ both have one row for each student. Also define a variable s_j that encapsulates the configuration of students to classrooms: For example, s_j tells how many students are in each classroom, and whether any two students are in the same classroom. $C(j)$ are the set of classrooms taught by teacher j, and $I(c)$ are the set of students in classroom i.

The best linear predictor of test scores given covariates and configuration is

$$E^*_{\mathcal{D}}\left[\boldsymbol{y}_j | \mu_j, \boldsymbol{x}_j, s_j\right] = \alpha + \mu_j + \boldsymbol{x}_j\boldsymbol{\beta}, \tag{1}$$

The teacher effect, μ_j, is teacher j's *value-added*, her causal effect on the outcome of

[5] I use bolded letters (i.e. $\boldsymbol{x}$) to represent vectors, and bolded and italicized letters (i.e. $\boldsymbol{x}$) to represent matrices.

[6] Maximum likelihood estimation is greatly simplified if $\bar{\bar{x}}_j$ is a precision-weighted mean, in a way that will be made clear.

interest. The best linear predictor of teacher effects given covariates is

$$E^*_{\mathcal{D}}\left[\mu_j|\bar{\bar{x}}_j, s_j\right] = \bar{\bar{x}}_j^T\boldsymbol{\lambda}. \tag{2}$$

$\boldsymbol{\lambda}$ is a vector governing the association of covariates with teacher quality. It could capture teacher-specific characteristics – for example, more experienced teachers are better – or reflect sorting – for example, teachers of honors classes may be better.

Combining Equations 1 and 2, $E^*_{\mathcal{D}}\left[\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\bar{x}}_j, s_j\right] = \alpha + \boldsymbol{x}_j\boldsymbol{\beta} + \bar{\bar{x}}_j^T\boldsymbol{\lambda}$. We can define errors $\tilde{\mu}_j$ and $\boldsymbol{\nu}_j$ with

$$\begin{aligned}
\mu_j &\equiv \bar{\bar{x}}_j\boldsymbol{\lambda} + \tilde{\mu}_j, \quad \tilde{\mu}_j \perp \bar{\bar{x}}_j \\
\boldsymbol{y}_j &\equiv \boldsymbol{x}_j\boldsymbol{\beta} + \bar{\bar{x}}_j^T\boldsymbol{\lambda} + \boldsymbol{\nu}_j, \quad \boldsymbol{\nu}_j \perp \boldsymbol{x}_j, \bar{\bar{x}}_j
\end{aligned} \tag{3}$$

In order to ascribe a casual interpretation to parameter estimates, we need sorting on observables. First, we need that variation in teacher effects that cannot be captured by covariates must be orthogonal to non-teacher shocks to test scores: $\tilde{\mu}_j \perp \left(\boldsymbol{\nu}_j - \tilde{\mu}_j\right)$. Second, more subtly, we need unobservable shocks to test scores to be *independent* of *assignments* to teachers: $\left(\boldsymbol{\nu}_j - \tilde{\mu}_j\right) \perp\!\!\!\perp s_j|\boldsymbol{x}_j, \bar{\bar{x}}_j$. To see why this second restriction is necessary, imagine that all teachers are identical – $\mu_j = 0 \quad \forall j$ – but some teachers are consistently assigned students with high values of $\boldsymbol{\nu}_j$. In that case, some teachers will consistently appear to have students that over- or under-perform what would be expected from their covariates, making it appear that teachers vary in their quality when they actually do not.

In order to make this model estimable via maximum likelihood, we need several more assumptions. First, $\boldsymbol{\beta}$ must correspond to an unrestricted linear predictor. That is, define the best linear predictor $\boldsymbol{\pi}$, so that

$$E^*_{\mathcal{D}}\left[\boldsymbol{y}_j|I_n \otimes \text{vec}(\boldsymbol{x}_j), \bar{\bar{x}}_j\right] = \left(I_n \otimes \text{vec}(\boldsymbol{x}_j)\right)\boldsymbol{\pi} + \bar{\bar{x}}\boldsymbol{\lambda}.$$

We need that $\left(I_n \otimes \text{vec}(\boldsymbol{x}_j)\right)\boldsymbol{\pi} = \boldsymbol{x}_j\boldsymbol{\beta}$. Finally, let's put more structure on the covariance of errors and assume homoskedasticity. Define $E\left[\boldsymbol{\nu}_j\boldsymbol{\nu}_j^T|s_j\right] \equiv \Sigma_j$. Denote parameters $\eta = \left(\alpha, \boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2_\mu, \sigma^2_\theta, \sigma^2_\varepsilon\right)$.

$$\begin{aligned}
\Sigma\left(\eta, \boldsymbol{x}_j, s_j\right)_{i,i'} &= \sigma^2_\mu + \sigma^2_\theta + \sigma^2_\varepsilon \quad \text{when } i = i' \\
\Sigma\left(\eta, \boldsymbol{x}_j, s_j\right)_{i,i'} &= \sigma^2_\mu + \sigma^2_\theta \quad \text{when } i \neq i' \text{ but } i \text{ and } i' \text{ are in the same class} \\
\Sigma\left(\eta, \boldsymbol{x}_j, s_j\right)_{i,i'} &= \sigma^2_\mu \quad \text{when } i \text{ and } i' \text{ are not in the same class}
\end{aligned}$$

$\text{Var}\left(\mu_j\right) = \text{Var}\left(\bar{\bar{x}}_j^T\boldsymbol{\lambda}\right) + \sigma^2_\mu$ is the amount of variance in y contributed by teachers; when teacher effects have a large variance, teachers are an important determinant of y. When σ^2_μ is large, there are large differences in teacher quality that are not predictable from observables. When variance in $\bar{\bar{x}}\boldsymbol{\lambda}$ is large, there are large differences in teacher quality that are predictable by observables. σ^2_θ and σ^2_ε are the shares of variance from classroom-level shocks and individual-specific shocks.

3.1 Maximum (Quasi-)Likelihood

No model like the one above has, to my knowledge, been estimated via maximum likelihood, but rather with GMM-like "moment-matching" procedures, as discussed at length below.

To generate a likelihood function, we must assume a functional form for the distributions of $\boldsymbol{y}_j$ and μ_j. Appendix A proves the validity of a quasi-likelihood interpretation: maximum likelihood based on normality delivers consistent estimates of η, even when the true distribution $\mathcal{D}$ does not have normal disturbances. Consider the model

$$\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\tilde{x}}_j, s_j \sim N\left(\boldsymbol{x}_j\boldsymbol{\beta} + \bar{\tilde{x}}_j^T\boldsymbol{\lambda}, \Sigma\left(\theta, s_j\right)\right)$$

with the corresponding likelihood function $f\left(\boldsymbol{y}_j, \boldsymbol{x}_j, \bar{\tilde{x}}_j, s_j; \theta\right)$. Appendix A proves that $\eta = \arg\max_\theta \mathbb{E}_{\mathcal{D}} \log f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\tilde{x}}_j, s_j; \theta\right)$. Appendix B derives a relatively simple closed-form solution for the likelihood, and defines classroom deviations from teacher means $\tilde{\bar{y}}_c = \bar{y}_c - \bar{\bar{y}}_{j(c)}$ and $\tilde{\bar{x}}_c = \bar{\boldsymbol{x}}_c - \bar{\bar{x}}_{j(c)}$.

Solving for the likelihood without integrating out teacher effects, as in Appendix Equation 26, gives an integral with a Bayesian interpretation that yields an Empirical Bayes posterior: Teacher effects are drawn $\mu_j \sim N\left(\bar{\tilde{x}}_j^T\boldsymbol{\lambda}, \sigma_\mu^2\right)$, and test scores are drawn $\bar{\bar{y}}_j \sim N\left(\mu_j + \bar{\tilde{x}}_j^T\boldsymbol{\beta}, \frac{1}{\sum_c h_c}\right)$, so the Empirical Bayes posterior of teacher j's value-added is

$$\mu_j \sim N\left(\frac{\sigma_\mu^2}{\sigma_\mu^2 + 1/\sum_{c\in C(j)} h_c}\left(\bar{\bar{y}}_j - \bar{\tilde{x}}_j^T\boldsymbol{\beta} - \alpha\right) + \frac{1/\sum_{c\in C(j)} h_c}{\sigma_\mu^2 + 1/\sum_{c\in C(j)} h_c}\bar{\tilde{x}}_j^T\boldsymbol{\lambda}, \left(\frac{1}{\sigma_\mu^2} + \sum_{c\in C(j)} h_c\right)^{-1}\right) \tag{4}$$

The solutions for $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\lambda}}$, and $\hat{\alpha}$ are intuitive. After concentrating out $\hat{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{\beta}}$ attempts to jointly minimize students' deviations from the classroom mean and classrooms' deviations from the teacher mean:

$$\hat{\boldsymbol{\beta}} = \arg\min_b \frac{1}{\hat{\sigma}_\varepsilon^2}\sum_i\left(\tilde{y}_i - \tilde{x}_i^T\boldsymbol{b}\right)^2 + \sum_j\sum_{c\in C(j)}\hat{h}_c\left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T\boldsymbol{b}\right)^2 \tag{5}$$

$\hat{\boldsymbol{\lambda}}$ and $\hat{\alpha}$ are given by weighted least squares, and minimize differences between teachers that can't be explained by differences within teachers:

$$\hat{\boldsymbol{\lambda}}, a = \arg\min_{\boldsymbol{\ell}, a}\sum_j\left(\frac{1}{\sum_c \hat{h}_c} + \hat{\sigma}_\mu^2\right)^{-1}\left(\bar{\bar{y}}_j - \bar{\tilde{x}}_j^T\hat{\boldsymbol{\beta}} - \bar{\tilde{x}}_j^T\boldsymbol{\ell} - a\right)^2 \tag{6}$$

When there are no classroom-level shocks, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}$ coincide with the estimands from a correlated random effects framework. When $\hat{\sigma}_\theta^2 = 0$, precisions h_c are proportional to the number of students in the class, so each observation is given equal weight. Equation 5 collapses to

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}}\sum_i\left(y_i - \bar{\bar{y}}_{j(i)} - \left(\boldsymbol{x}_i - \bar{\tilde{x}}_{j(i)}\right)^T\boldsymbol{b}\right)^2,$$

and Equation 6 becomes

$$\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\lambda}}, \hat{\alpha} = \arg\min_{\boldsymbol{c},a} \sum_j \left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T \boldsymbol{c} - a \right)^2.$$

These equations yield the same coefficients as running the regression (Chamberlain (1984), Chamberlain (1982))

$$y_i = \alpha + x_i^T \boldsymbol{\beta} + \bar{\bar{x}}_{j(i)}^T \boldsymbol{\lambda} + \varepsilon_i.$$

3.1.1 Inference

Since this is MLE, asymptotic inference is (conceptually) easy. Appendix D gives the first derivative of the likelihood function. When comparing estimators on real datasets in Section 5, I report bootstrap confidence intervals and p-values.

3.1.2 Bias correction to variance of teacher effects

The quantity of interest is

$$\text{Var}(\mu_j) = \text{Var}\left(\bar{\bar{x}}_j^T \boldsymbol{\lambda} \right) + \sigma_\mu^2$$

An obvious estimator is the sample analog:

$$\widehat{\text{Var}}(\mu|\hat{\boldsymbol{\lambda}}) = \frac{1}{n} \sum_j \left(\bar{\bar{x}}_j^T \hat{\boldsymbol{\lambda}} \right)^2 - \left(\frac{1}{n} \sum_j \bar{\bar{x}}_j^T \hat{\boldsymbol{\lambda}} \right)^2 + \hat{\sigma}_\mu^2$$

However, the sample analog is biased upwards. $\mathbb{E}\left[\text{Var}\left(\bar{\bar{x}}_j^T \hat{\boldsymbol{\lambda}} | \hat{\boldsymbol{\lambda}} \right) \right] > \mathbb{E}\left[\text{Var}\left(\bar{\bar{x}}_j^T \boldsymbol{\lambda} \right) \right]$, for a clear reason: estimation error in $\hat{\boldsymbol{\lambda}}$ will tend to make this quantity larger. Imagine that $\boldsymbol{\lambda} = 0$: $\hat{\boldsymbol{\lambda}}$ will not be zero, so there will appear to be some correlation between teacher effects and covariates when there is not. Specifically, as shown in Appendix Proof 27, the sample analog is biased upwards by

$$\mathbb{E}\left[\text{Var}\left(\bar{\bar{x}}^T \hat{\boldsymbol{\lambda}} | \hat{\boldsymbol{\lambda}} \right) \right] - \text{Var}\left(\bar{\bar{x}}^T \boldsymbol{\lambda} \right) = \mathbb{E}\left[\left(\bar{\bar{x}} - \mathbb{E}\, \bar{\bar{x}} \right)^T \text{Cov}\left(\hat{\boldsymbol{\lambda}} \right) \left(\bar{\bar{x}} - \mathbb{E}\, \bar{\bar{x}} \right) \right]. \tag{7}$$

Therefore, a bias-corrected estimator of the variance of teacher effects is

$$\widehat{\text{Var}}\left(\mu_j \right) = \overbrace{\frac{1}{n} \sum_j \left(\bar{\bar{x}}_j^T \boldsymbol{\lambda} \right)^2 - \left(\frac{1}{n} \sum_j \bar{\bar{x}}_j^T \boldsymbol{\lambda} \right)^2}^{\text{predictable variance}} + \underbrace{\hat{\sigma}_\mu^2}_{\text{unpredictable variance}} - \overbrace{\frac{1}{n} \sum_j \left(\bar{\bar{x}}_j^T - \frac{1}{J} \sum_k \bar{\bar{x}}_k \right)^T \hat{\Sigma}_\lambda \left(\bar{\bar{x}}_j - \frac{1}{J} \sum_k \bar{\bar{x}}_k \right)}^{\text{bias correction}}. \tag{8}$$

where $\hat{\Sigma}_\lambda$ is the asymptotic variance of $\hat{\boldsymbol{\lambda}}$.

3.2 Empirical Bayes estimator from Kane and Staiger (2008)

Kane and Staiger (2008) develop a model that other value-added papers use as a baseline, such as Chetty *et al.* (2014). As discussed in Section 2, Kane and Staiger (2008) experimentally validated value-added scores and did not reject the hypothesis that the scores were forecast-unbiased. However, their estimates also suggest that, when controlling for student fixed effects, the value-added scores actually understate the magnitude of teacher effects.

Guarino *et al.* (2014) and others note that this estimator is not consistent when teacher effects are correlated with covariates. Section 3.2.4 shows that although the estimator is consistent when teacher effects are not correlated with covariates – $\boldsymbol{\lambda} = 0$ –, this estimator is asymptotically downward biased when $\boldsymbol{\lambda} \neq 0$. 3.3 lays out a "modified-KS" estimator that amends estimation to follow a fixed effects rather than random effects assumption, and consistently estimates the model laid out in Section 3.1 under only a sorting on observables assumption.

3.2.1 Estimation

Kane and Staiger (2008)'s estimation procedure, like many other Empirical Bayes procedures and like the maximum likelihood procedure above, proceeds in two phases. First, we estimate the parameters of the model: $\boldsymbol{\beta}$, σ^2_μ, σ^2_θ, and σ^2_ϵ. Then we estimate each teacher's value of μ_j using the distribution described by the previously-estimated parameters as a prior.

The first stage, estimation of parameters, itself comprises two steps. First, we estimate $\hat{\boldsymbol{\beta}}$, then we use $\hat{\boldsymbol{\beta}}$ to generate residuals. Next, we use a "moment-matching" procedure to estimate the variances σ^2_μ, σ^2_θ, and σ^2_ϵ based on variances and covariances of residuals. In more detail:

$\hat{\boldsymbol{\beta}}$ is estimated by regressing outcomes y_i on covariates $\boldsymbol{x}_i$. This gives a consistent and unbiased estimate of $\boldsymbol{\beta}$ if and only if teacher effects are uncorrelated with covariates; otherwise, this estimate will suffer from omitted variable bias:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\min_b \sum_i \left(y_i - x_i^T \boldsymbol{b}\right)^2 \\
&= \boldsymbol{\beta} + \left(\sum \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \sum \boldsymbol{x}_i \left(\mu_{j(i)} + \nu_i\right) \\
\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= \boldsymbol{\beta} + \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i \mu_{j(i)}\right]
\end{aligned}$$

(The modified-KS estimator presented in the next section explores estimating $\boldsymbol{\beta}$ using within-teacher variation, which corrects this omitted variable bias.)

In order to estimate σ^2_μ, use the following procedure. Let $C(j)$ denote the set of classes taught by teacher j. σ^2_μ is the average product of mean residuals in pairs of classes taught by the same teacher: [7]

[7] Kane and Staiger use a different version of this procedure, using only sequential pairs of classrooms. (Check this!)

$$\hat{\sigma}_\mu^2 = \frac{2}{\sum_j |C(j)| \left(|C(j)| - 1\right)} \sum_j \sum_{c,c' \in C(j): c \neq c'} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}}\right) \tag{9}$$

To estimate $\hat{\sigma}_\theta^2$ and $\hat{\sigma}_\varepsilon^2$, we use similar "moment-matching" ideas. Since ε is responsible for within-classroom variation in $\tilde{y}$, σ_ε^2 is the mean variance of $\tilde{y}_i$ within a classroom:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N_{\text{students}} - N_{\text{classes}}} \sum_j \left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}\right)^2$$

$\hat{\sigma}_\theta^2$ is chosen to explain the variance in y_i that is not explained by μ, ε, or $\hat{\boldsymbol{\beta}}$:

$$\hat{\sigma}_\theta^2 = \widehat{\text{Var}}(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) - \hat{\sigma}_\mu^2 - \hat{\sigma}_\varepsilon^2.$$

3.2.2 Inference

We can reformulate this "moment-matching" procedure as the solution to a set of moment functions. After setting up a moment function, we can estimate the asymptotic distribution of the parameters either through the Bayesian Bootstrap, as in Chamberlain (2013), or through the Generalized Method of Moments.

Denote the parameters $\eta = (\boldsymbol{\beta}, \sigma_\mu^2, \sigma_\varepsilon^2, \sigma_\theta^2)$. Letting $S(j)$ be the set of students with teacher j, $S(j) = \{i : j(i) = j\}$, the moment function, which is at the teacher level, is

$$g_j(\eta) = \begin{pmatrix} \sum_{i \in S(j)} x_i (y_i - x_i^T \hat{\boldsymbol{\beta}}) \\ \sum_{c,c' \in C(j)} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}}\right) - \hat{\sigma}_\mu^2 \left(\frac{|C(j)|(|C(j)|-1)}{2}\right) \\ \sum_{i \in S(j)} \left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}\right)^2 - \hat{\sigma}_\varepsilon^2 \left(|S(j)| - |C(j)|\right) \\ \sum_{i \in S(j)} \left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}\right)^2 - |S(j)| \left(\hat{\sigma}_\mu^2 - \hat{\sigma}_\theta^2 - \hat{\sigma}_\varepsilon^2\right) \end{pmatrix}$$

To take the n^{th} Bayesian Bootstrap draw, draw weights $\omega^n \in \mathbb{R}^{\text{Nteachers}}$ according to $\omega^n \sim \text{Dirichlet}\,(1, 1, \ldots, 1)$ (Rubin, 1981). Bootstrap draws of parameters become

$$\begin{aligned}
\hat{\beta}^n &= \left(\sum_i \omega_{j(i)}^n x_i x_i^T\right)^{-1} \sum_i \omega_{j(i)}^n x_i y_i \\
\hat{\sigma}_\mu^{2(n)} &= \frac{2}{\sum_j \omega_j^n |C(j)| \left(|C(j)| - 1\right)} \sum_j \sum_{c,c' \in C(j)} \omega_j^n \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}}^n\right) \\
\hat{\sigma}_\varepsilon^{2(n)} &= \frac{1}{\sum_j \omega_j^n \left(|S(j)| - |C(j)|\right)} \sum_i \omega_{j(i)}^n \left(\tilde{y}_i - \tilde{x}_i^T \hat{\beta}^n\right)^2 \\
\hat{\sigma}_\theta^{2(n)} &= \frac{1}{\sum_j \omega_j^n |S(j)|} \sum_i \omega_{j(i)}^n \left(y_i - x_i^T \hat{\beta}^n\right)^2 - \hat{\sigma}_\mu^{2(n)} - \hat{\sigma}_\varepsilon^{2(n)}
\end{aligned}$$

3.2.3 Individual Teacher Effects

While $\bar{\tilde{y}}_j - \bar{\tilde{x}}_j^T \hat{\beta}$ is an unbiased estimate of μ_j, Kane and Staiger use shrinkage to produce a best linear predictor of μ_j. First, generate the precision $h_c = \text{Var}(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \boldsymbol{\beta})^{-1}$ of each mean

classroom residual; these are the same precisions used for maximum likelihood in Equation 24. Then construct a precision-weighted mean using h_c and multiply it by shrinkage factor ρ_j use linear shrinkage ρ_j and precisions h_c to generate a mean squared error-minimizing estimate of μ_j:

$$\begin{aligned}\hat{\mu}_j &= \rho_j \frac{\sum_c h_{c:j(c)=j}\left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \boldsymbol{\beta}\right)}{\sum_c h_c} \\ \rho_j &= \arg\min_{\rho} \mathbb{E}\left[\left(\rho \frac{\sum_c h_c \tilde{y}_c}{\sum_c h_c} - \mu_j\right)^2\right] = \frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2 + \frac{1}{\sum_c h_c}}\end{aligned} \tag{10}$$

Kane and Staiger note that when μ, θ, and ε are normally distributed, Equation 10 has a Bayesian interpretation. The estimated variances $\hat{\sigma}_\mu^2$, $\hat{\sigma}_\theta^2$, and $\hat{\sigma}_\varepsilon^2$ are treated as a prior and observed test scores as data to create Empirical Bayes maximum a posteriori estimates of teacher effects, which shrink mean residuals towards zero.

Equation 10 equals Equation 4, from maximum likelihood, when $\hat{\boldsymbol{\lambda}} = 0$: Conditional on parameter estimates, both procedures deliver the same estimated individual teacher effects. However, even with the imposition of $\boldsymbol{\lambda} = 0$, the procedures will generally not estimate the same parameters. When estimating $\hat{\beta}$, the Kane and Staiger procedure implicitly gives each observation equal weight, while maximum likelihood uses precision weighting to put relatively more weight on smaller classes. Maximum likelihood uses precision weights to give relatively Even with the imposition of $\boldsymbol{\lambda} = 0$, the estimates of $\hat{\beta}$ will not generally be the same unless σ_θ^2.

3.2.4 Inconsistency and bias under misspecification

Consistency and bias of $\hat{\sigma}_\mu^2$

As discussed extensively in Guarino *et al.* (2014) and mentioned in Chetty *et al.* (2014), the Kane and Staiger estimator will only be valid if there is no correlation between observable student characteristics and teacher value-added. Their work demonstrates that $\hat{\boldsymbol{\beta}}$ will be biased when estimated in a regression that omits teacher fixed effects; this project demonstrates that omitted variable bias in $\hat{\boldsymbol{\beta}}$ leads to an asymptotic *negative* bias in $\hat{\sigma}_\mu^2$. Intuitively, variation in teacher effects that is correlated with student characteristics is incorrectly attributed to the student characteristics.

Consistency of $\hat{\boldsymbol{\beta}}$, independence of errors between teachers, and finite second moments are jointly sufficient for consistency of $\hat{\sigma}_\mu^2$. Unbiasedness of $\hat{\boldsymbol{\beta}}$ is necessary for unbiasedness of $\hat{\sigma}_\mu^2$.

Since we know the probability limit of $\hat{\boldsymbol{\beta}}$, and $\hat{\sigma}_\mu^2$ is a smooth function of $\hat{\boldsymbol{\beta}}$, we also know the probability limit of $\hat{\sigma}_\mu^2$. Note that Equation 9 gives

$$\begin{aligned}\hat{\sigma}_\mu^2 = \frac{1}{N_{\text{teachers}}} \sum_j \frac{2}{|C(j)|\,|C(j)-1|} \sum_{c,c' \in C(j)} &\left(\mu_j + \theta_c + \bar{\varepsilon}_c + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \bar{\boldsymbol{x}}_c\right) \\ &\left(\mu_j + \theta_{c'} + \bar{\varepsilon}_{c'} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \bar{\boldsymbol{x}}_{c'}\right).\end{aligned}$$

Also,

$$\begin{aligned}\boldsymbol{\beta} \rightarrow_p &\boldsymbol{\beta} + \mathbb{E}\left[x_i x_i^T\right]^{-1} \mathbb{E}\left[x_i \mu_{j(i)}\right] \\ =&\boldsymbol{\beta} + \mathbb{E}\left[x_i x_i^T\right]^{-1} \mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T \boldsymbol{\lambda}\right]\end{aligned}$$

Therefore, letting c and c' represent a randomly chosen pair of classes taught by the same teacher, $\hat{\sigma}_\mu^2$ converges to

$$\begin{aligned}\hat{\sigma}_\mu^2 \rightarrow_p &\mathbb{E}\left[\mu_j^2\right] - 2\,\mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right]\boldsymbol{\lambda} + \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \bar{x}_c \bar{x}_c^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] \\ &= \mathrm{Var}\left(\mu_j\right) + \boldsymbol{\lambda}^T \mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right]\left(\mathbb{E}\left[x_i x_i^T\right]^{-1} \mathbb{E}\left[\bar{x}_c \bar{x}_{c'}^T\right] \mathbb{E}\left[x_i x_i^T\right]^{-1} - 2\,\mathbb{E}\left[x_i x_i^T\right]^{-1}\right) \mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right] \boldsymbol{\lambda}\end{aligned} \tag{11}$$

Equation 11 shows that if there is no sorting on observables (so $\boldsymbol{\lambda} = 0$), then $\hat{\sigma}_\mu^2 \rightarrow_p \sigma_\mu^2$. Equation 11 also allows us to bound the asymptotic bias by considering two extremes: the case in which average student characteristics are perfectly correlated across different classrooms taught by the same teacher, so that $\mathbb{E}\left[\bar{x}_c \bar{x}_{c'}^T\right] = \mathbb{E}\left[\bar{x}_c \bar{x}_{c'}^T\right]$, and the case in which average characteristics are perfectly anti-correlated. Appendix E shows that

$$-b^T\,\mathbb{E}\left[x_i x_i^T\right] b \leq b\,\mathbb{E}\left[\bar{x}_c \bar{x}_{c'}^T\right] b \leq b^T\,\mathbb{E}\left[x_i x_i^T\right] b \quad \forall b \in \mathcal{R}^k.$$

When the first constraint binds, students are identical within a classroom, and the characteristics of two classrooms taught by the same teacher are perfectly anti-correlated. When the second constraint binds, all students taught by the same teacher are identical. Plugging these bounds into Equation 11,

$$-3 \leq \frac{\mathrm{plim}\left(\hat{\sigma}_\mu^2\right) - \mathrm{Var}\left(\mu_j\right)}{\boldsymbol{\lambda}^T\,\mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right]\mathbb{E}\left[x_i x_i^T\right]^{-1}\mathbb{E}\left[\bar{\bar{x}}_j \bar{\bar{x}}_j^T\right]\lambda} \leq -1 \tag{12}$$

Equation 12 makes several facts apparent. The estimator is always negatively biased in asymptopia, and bias is more severe when sorting is strong. This happens when$\boldsymbol{\lambda}$ is large in magnitude. The asymptotic bias becomes smaller when classrooms taught by the same teacher are very similar to each other.

3.3 Modified version of above estimator: Modified-KS

As discussed above, omitted variables bias parameter estimates in the Kane and Staiger estimator. Chetty *et al.* (2014) suggest remedying this by including teacher fixed effects when residualizing. That is, we obtain $\hat{\boldsymbol{\beta}}$ as the coefficient on x_i in a regression of outcomes on x_i and teacher fixed effects [8]. In a similar spirit, Guarino *et al.* (2014) discuss a similar issue in the context of a slightly different value-added procedure from that of Kane and Staiger

[8]Chetty *et al.* (2014) use an estimator much more complicated than the Kane and Staiger estimator; they model the "drift" in teacher value-added across years. In this section, I use their modification to the estimation of $\hat{\boldsymbol{\beta}}$ but do not study the rest of their model.

(2008): the “mixed model" of Ballou *et al.* (2004), which differs from the model of Kane and Staiger (2008) in that it does not explicitly model classroom effects (θ_c). This model assumes that teacher effects are uncorrelated with student covariates, and performs poorly when that assumption is false. Guarino *et al.* (2014) explain that “estimators that include the teacher assignment indicators along with the covariates in a multiple regression analysis" perform better. In this section and in the simulations in Section 4, I show that using within-teacher variation to estimate $\hat{\boldsymbol{\beta}}$ improves the *asymptotic* performance of the estimator when teacher effects are correlated with observables, but still leaves a bias of ambiguous sign in finite samples, and Monte Carlo exercises suggest that this estimator may be more variable than the Kane and Staiger estimator.

Equation 11 makes the improved asymptotic performance of the estimator clear: This procedure yields a consistent estimate of $\boldsymbol{\beta}$, and a consistent estimate of $\hat{\boldsymbol{\beta}}$ leads to yields a consistent estimate of $\hat{\sigma}^2_\mu$. However, $\hat{\sigma}^2_\mu$ is still biased in finite samples:

$$\begin{aligned}\mathbb{E}\,\widehat{\mathrm{Var}}(\mu_j) - \mathrm{Var}(\mu_j) =& \frac{1}{N_{\text{teachers}}} \sum_j \frac{2}{|C(j)|\,|C(j)-1|} \sum_{c,c' \in C(j)} \mathbb{E}\left[\left(\hat{\beta}-\beta\right)^T \bar{x}_c \bar{x}_{c'}^T \left(\hat{\beta}-\beta\right)\right] \\ & - \frac{2}{N_{\text{teachers}}} \sum_j \frac{2}{|C(j)|\,|C(j)-1|} \sum_{c,c' \in C(j)} \mathbb{E}\left[\nu_c \left(\hat{\beta}-\beta\right)^T \right] \bar{x}_{c'}\end{aligned}$$

The sign of the bias is in general ambiguous, but seems to be positive in simulations, both with simulated outcomes and with permutation tests. The first term will be positive if student characteristics are sufficiently correlated across classrooms taught by the same teacher. This term will disappear quickly as the number of teachers increases.

3.3.1 Inference

Inference is the same as in the Kane and Staiger estimator, except that the first component of the moment condition changes to reflect that $\hat{\boldsymbol{\beta}}$ is now estimated off of within-teacher variation.

$$g_j(\eta) = \begin{pmatrix} \sum_{i \in S(j)} \left(x_i - \frac{1}{|S(j)|} \sum_{i' \in S(j)} \right) \left(y_i - x_i^T \hat{\boldsymbol{\beta}} - \frac{1}{|S(j)|} \sum_{i' \in S(j)} \left(y_{i'} - \boldsymbol{x}_{i'}^T \hat{\boldsymbol{\beta}} \right) \right) \\ \sum_{c,c' \in C(j)} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}} \right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}} \right) - \hat{\sigma}^2_\mu \left(\frac{|C(j)|(|C(j)|-1)}{2} \right) \\ \sum_{i:j(i)=j} \left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}} \right)^2 - \hat{\sigma}^2_\varepsilon \left(|S(j)| - |C(j)| \right) \\ \sum_{i:j(i)=j} \left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} \right)^2 - |S(j)| \left(\hat{\sigma}^2_\mu - \hat{\sigma}^2_\theta - \hat{\sigma}^2_\varepsilon \right) \end{pmatrix}$$

4 Simulation Experiments

In order to assess the validity of these estimators with data as realistic as possible, I use the covariates and assignment structure of a real dataset, but simulated individual effects and outcomes. Thus the variance of individual effects and their covariance with covariates is known. I use simulations based on the normal distribution, the t distribution with three degrees of freedom, and the Poisson distribution.

The dataset used is a set of assignments of elite bureaucrats in the Indian Administrative Service to districts in India. This data is described in much more detail in Chapter 3

of this dissertation. Bureaucrats are assigned to a state at the beginning of their careers, and afterwards are quasi-randomly assigned to different districts in the same state every one to two years for several years. They will often serve several postings as powerful "District Collectors," responsible for various tasks in administering the district. Table 3 shows summary statistics. This data is small compared to educational datasets, with 90,138 district-month observations, 2,965 District Collectors, and 5,048 collector-district pairs. The second and third columns of Table 3 restrict the sample to observations where an outcome (night light intensity or project completions) is observed and to the largest "connected set" where the bipartite graph of people and districts is connected by bureaucrat transfers [9]. In this section, I use the full sample since outcomes are simulated, and in the next section, I show results on real data.

Table 3: *Summary statistics for whole sample and largest connected set*

	Full		Largest conn. set (lights)		Largest conn. set (projects)	
	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev
Postings per collector	1.70	0.99	1.78	1.03	1.79	1.04
Posting length (months)	17.86	11.83	14.78	9.81	17.19	11.25
District-month observations	90,138		62,488		74,013	
Postings	5,048		4,229		4,305	
Collectors	2,965		2,376		2,410	
Districts	529		424		428	
States	34		20		22	
Start year	1996		1996		1996	
End year	2014		2013		2014	

Notes. Full sample consists of all district collectors with a district assignment and serving after 1996. The "largest connected set" for an outcome refers to the subset of districts connected by district collector mobility where the outcome is not null. Number of district postings per collector is defined as the number of distinct districts a collector has been assigned to over his or her career. Tenure length is based off the begin and end dates of a posting in each officer's record sheet.

Bureaucrat effects have a variance of 1, and the outcome has a variance of 100. I use controls $x_i = (x_i^0, x_i^1, \ldots, x_i^{12})$, where x_i^0 is the number of months since January 1996 and $x_i^1, \ldots, x_i^{12}$ are month-of-year fixed effects to account for seasonality. I simulate bureaucrat effects and outcomes for various values of ρ, the correlation between bureaucrat effects and time:

$$\mu_j \sim N\left(\rho \frac{\bar{\bar{x}}_j^0}{\sqrt{\text{Var}(\bar{\bar{x}}_j^0)}}, 1-\rho^2\right)$$
$$y_i \sim N(\mu_{j(i)}, 99)$$

[9] Restricting to the one connected set is necessary in specifications that include district fixed effects, because otherwise bureaucrat effects and district effects cannot both be identified.

Figure 1: *Results from 1000 Monte Carlo simulations illustrate that, as expected, the Kane and Staiger estimator is biased downwards when teacher effects are correlated with student characteristics, and that maximum likelihood estimates are biased upwards regardless of the correlation between teacher effects and student characteristics. Bias-corrected maximum likelihood estimates are close to unbiased. Error bars plot 95% confidence intervals based on the standard deviation of the bias.*

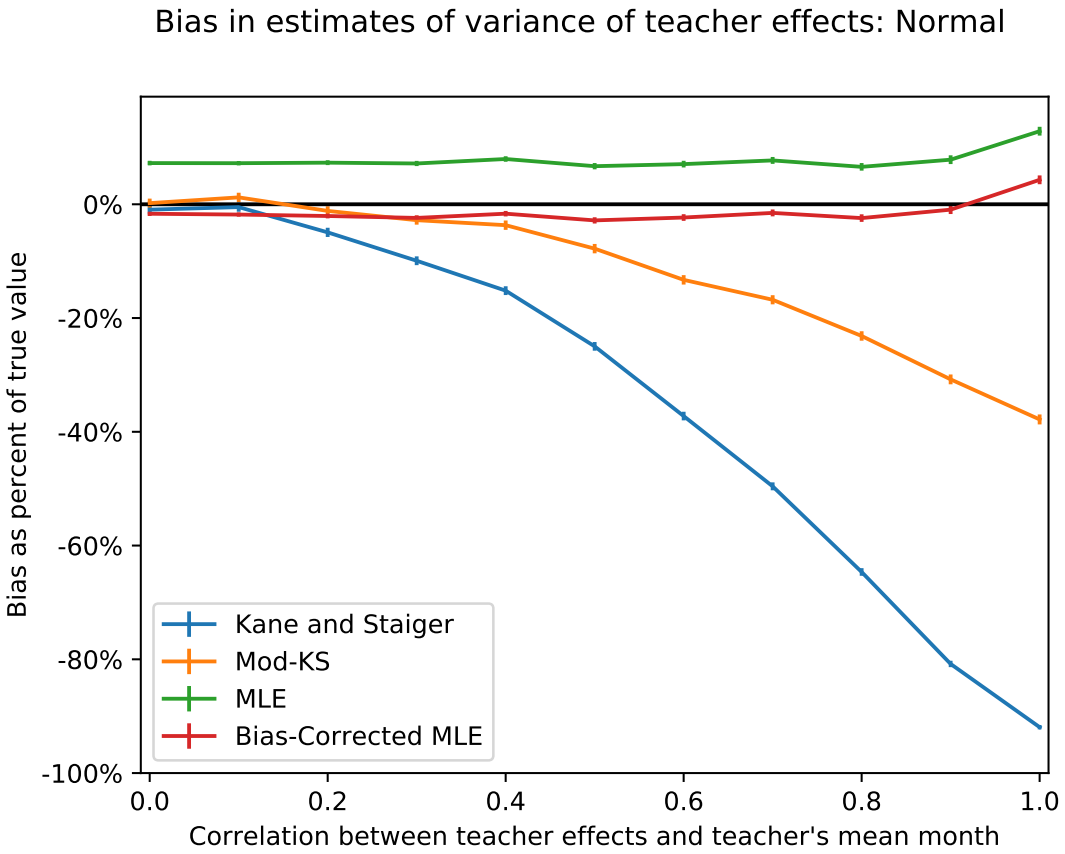


Table 4: *Bias and MSE for each estimator, when teacher effects are correlated with month at* $\rho = 0$, 0.5, *and* 1.

	Bias			Sqrt MSE		
Estimator	0.0	0.5	0.9	0.0	0.5	0.9
Kane and Staiger	-0.9%	-25.0%	-80.8%	27.7%	36.4%	83.1%
	(0.8%)	(0.8%)	(0.6%)	(2.1%)	(1.8%)	(0.6%)
Mod-KS	0.2%	-7.8%	-30.8%	27.7%	29.7%	43.0%
	(0.8%)	(0.8%)	(0.9%)	(2.2%)	(2.2%)	(1.6%)
MLE	7.2%	6.7%	7.8%	15.8%	20.3%	27.9%
	(0.4%)	(0.6%)	(0.8%)	(2.1%)	(2.4%)	(2.4%)
Bias-Corrected MLE	-1.7%	-2.8%	-0.9%	13.7%	19.3%	26.8%
	(0.4%)	(0.5%)	(0.8%)	(2.2%)	(2.1%)	(2.2%)

Results from simulations accord with predictions from theory: the Kane and Staiger estimator is increasingly downward-biased as ρ increases; maximum likelihood is biased upwards; and the bias correction of Equation 8 greatly improves the bias of the maximum likelihood estimator. Figure 1 plots the mean bias over 1,000 Monte Carlo simulations for each of eleven values of ρ, and Table 4 gives the bias and mean squared error of each estimator.

Figure **??** shows the variance and squared bias for each estimator, summing to mean squared error. It shows that likelihood-based estimators have lower variance than moment matching-based estimators, at least in this simulation exercise.

To investigate whether maximum likelihood still performs well when misspecified, I posit that outcomes are generated according to either a Poisson process, leading to skewed, discrete outcomes, or according to the t distribution, which is much heavier-tailed than the normal distribution. In both of these simulations, I maintain that bureaucrats account for 1% of variance and are correlated with a covariate at between 0 and 1, and I use the same controls as in the normal model.

Poisson-distributed outcomes have mean 0.99 and variance 1. Bureaucrat effects and outcomes are generated according to

$$\begin{aligned}
\tilde{\mu}_j &\sim N\left(\frac{0.1\rho}{\sqrt{\text{Var}\,\bar{\bar{x}}_j}}(\bar{\bar{x}}_j - \text{mean}(\bar{\bar{x}}_j), \quad 0.01(1-\rho^2)\right) \\
y_i &\sim \text{Poisson}\left(0.99 + \tilde{\mu}_j\right).
\end{aligned}$$

In the t distribution, bureaucrat effects have variance 1 and outcomes have variance 100:

$$\begin{aligned}
\mu_j &\sim N\left(\frac{\rho}{\sqrt{\text{Var}\,\bar{\bar{x}}_j}}\bar{\bar{x}}_j, \quad 1-\rho^2\right) \\
\epsilon_i &\sim t(3) \\
y_i &= \mu_{j(i)} + \sqrt{33}\epsilon_i
\end{aligned}$$

Table 5: *Bias and MSE for each estimator with Poisson-distributed outcomes, when teacher effects are correlated with month at* $\rho = 0$, 0.5, *and* 1.

	Bias			Sqrt MSE		
Estimator	0.0	0.5	0.9	0.0	0.5	0.9
Kane and Staiger	-1.4%	-24.7%	-78.8%	28.3%	37.4%	81.7%
	(1.0%)	(1.0%)	(0.7%)	(2.4%)	(2.2%)	(0.8%)
Mod-KS	-0.2%	-6.7%	-29.5%	28.3%	30.1%	43.7%
	(1.0%)	(1.0%)	(1.1%)	(2.4%)	(2.3%)	(2.0%)
MLE	7.1%	8.7%	8.0%	16.1%	20.9%	28.5%
	(0.5%)	(0.6%)	(0.9%)	(2.3%)	(2.4%)	(2.7%)
Bias-Corrected MLE	-1.8%	-1.6%	-2.1%	14.2%	19.0%	27.4%
	(0.5%)	(0.6%)	(0.9%)	(2.5%)	(2.3%)	(2.4%)

Table 6: *Bias and MSE for each estimator with t-distributed outcomes, when teacher effects are correlated with month at* $\rho = 0, 0.5,$ *and* $1.$

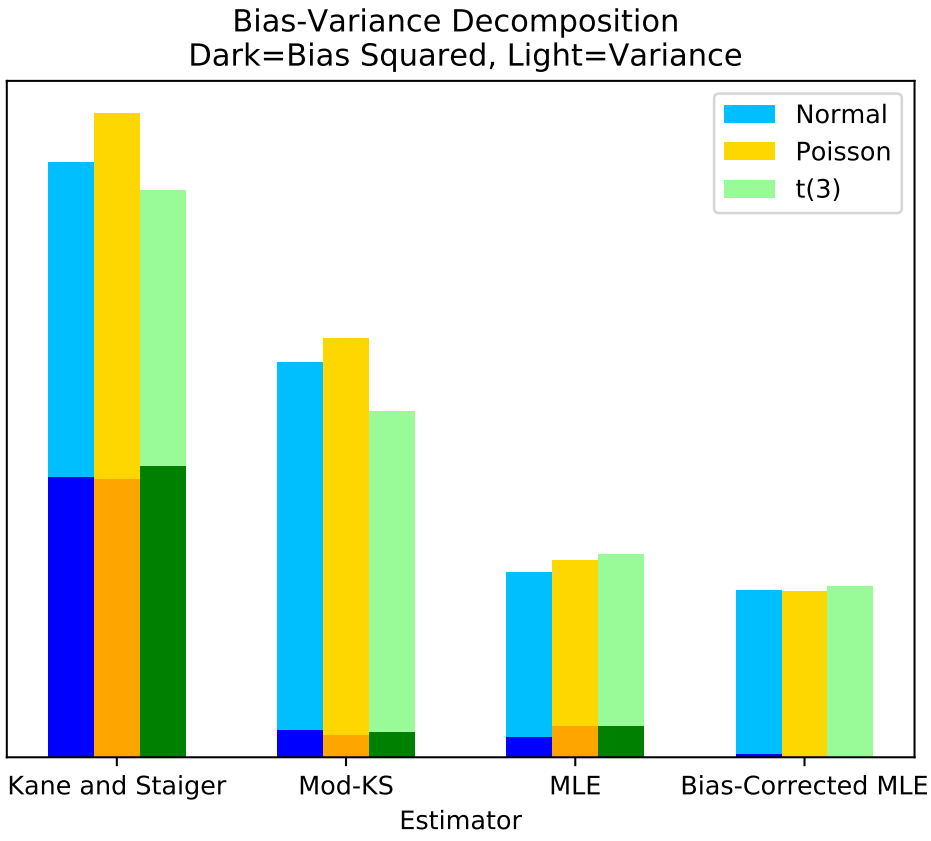
	Bias			Sqrt MSE		
Estimator	0.0	0.5	0.9	0.0	0.5	0.9
Kane and Staiger	-1.5%	-26.6%	-80.8%	27.1%	36.6%	82.7%
	(1.5%)	(1.4%)	(1.0%)	(4.7%)	(3.4%)	(1.1%)
Mod-KS	-0.2%	-8.8%	-30.3%	27.4%	29.0%	42.1%
	(1.5%)	(1.6%)	(1.6%)	(4.6%)	(4.3%)	(3.4%)
MLE	7.2%	7.7%	7.6%	17.8%	21.0%	26.9%
	(0.9%)	(1.1%)	(1.4%)	(4.3%)	(4.3%)	(4.5%)
Bias-Corrected MLE	-1.7%	-1.8%	-1.1%	16.0%	19.6%	25.7%
	(0.9%)	(1.1%)	(1.4%)	(4.1%)	(4.1%)	(4.1%)

Tables 5 and 6, corresponding to 4 for the normal distribution, show the bias and square root of mean squared error, as a percentage of the true value, for Poisson-distributed and t-distributed data, respectively. They show results from $\rho = .5$.

Appendix Figures 3 and 4 plot bias as a function of ρ for each estimator. They look extremely similar to figure 1, the analogous figure for the normal distribution.

Figure **??** shows a bias-variance decomposition for each estimator and for each distribution, at $\rho = .5$. The dark parts of the bars plot bias squared and the light parts plot variance, so the height of the bar represents mean squared error. Given the substantial correlation between bureaucrat effects and covariates, the Kane and Staiger estimator performs very poorly due to it substantial bias. Modified-KS and MLE are both much less biased. Maximum quasi-likelihood has a lower variance that Modified-KS, perhaps because it is able to use information from bureaucrats who only serve in one district, and bias-corrected nearly eliminates the bias. These patterns hold for every distribution, although the t distribution seems to generate higher mean squared error.

Figure 2: *Mean variance and mean squared bias (adding up to mean squared error) for each estimator at* $\rho = 0.5$*, from 1000 Monte Carlo simulations.*



5 Comparison on real data

In this section, I compare the results of each estimator on two different data sets. Consistent with predictions, Kane and Staiger gives the smallest estimates, and maximum likelihood the largest.

The outcome variable in Table 7 is the number of project completions listed by the Center for Monitoring the Indian Economy's CapEx dataset. The CapEx data details over 45,000 projects begun since 1996 which create new productive capacity and aims to capture all projects with a capital cost of over 10 million rupees, or $154,000 US dollars. In estimating value-added models, I control for a linear time trend and for the districts mean project completions for each of its three previous District Collectors.

Table 7 shows that, depending on estimator, variation in bureaucrat quality accounts for 0.038% to 0.741% of variance in project completions. Given the very large size of these projects, and the fact that District Collectors perform many activities that would not influence project completions, 0.7% would be a very large result, implying that District Collectors quality might be an important determinant of economic activity. However, a bootstrap p-value (Andrews, 2000) shows that the results are in fact insignificant.

Table 9 shows estimates from a dataset of eighth grade math teachers in New York City, with eighth grade math test scores as the dependent variable. Table 8 gives summary statistics. This dataset is described in far more detail in the third chapter of this dissertation.

Table 9 shows a more consistent and less biased picture than the India bureaucrat results: point estimates of the variance of teacher quality are 3.1% to 4.2% of the variance in test scores, falling to 3.1% to 3.9% after a bootstrap bias correction Horowitz (2001). Bootstrap 95% confidence intervals show that the likelihood-based estimators are slightly more dispersed than the moment-matching estimators. A bootstrap bias correction does not affect the Kane and Staiger or Mod-KS estimates, while it does reduce the MLE and MLE Bias-Corrected estimates each by about 0.3 percentage points.

Table 7: *IAS bureaucrats: Comparison of point estimates from different estimators, with a p-value from 1000 bootstrap iterations.*

	Kane and Staiger	Mod-KS	MLE	MLE Bias-Corrected
Point Est.	0.038%	0.270%	0.741%	0.719%
Boot p	0.379	0.246	0.397	0.400

Table 8: *New York City students of eighth grade math teachers: Summary statistics.*

index	Mean	St. Dev	Min	Max	Missing
Grade	8	0	8	8	0%
Year	2008.9	2.02	2006	2013	0%
Disabled	0.13	0.34	0	1	0%
Female	0.49	0.5	0	1	0%
English Language Learner	0.1	0.3	0	1	0%
Free Lunch	0.81	0.39	0	1	0%
Days absent	14.09	14.3	0	179	0%
Days present	167.1	15.2	2	186	0%
Days Absent Lag (Z-Score)	0.1	0.78	-13.66	1.11	2.79%
Math Score (Z-Score)	0.13	0.96	-6.35	3.78	0%
Math score lag (Z-Score)	0.11	0.96	-5.94	3.92	2.76%
ELA Score (Z-Score)	0.07	0.98	-10.67	6.56	0%
ELA Score Lag (Z-Score)	0.09	0.96	-11.1	6.63	5.46%
4-Year Graduation	0.68	0.47	0	1	31.44%
4-Year Graduation, Regents Diploma	0.44	0.5	0	1	31.43%
4-Year Graduation, Advanced Regents Diploma	0.2	0.4	0	1	31.39%
N = 337,070					

Table 9: *Comparison of point estimates from different estimators, with a p-value from 1000 bootstrap iterations and 95% bootstrap confidence interval.*

	Point Est.	Percentile CI	Boot Bias Corr. Est.	Boot p
Kane and Staiger	3.117%	(2.904%, 3.327%)	3.133%	0.000
Mod-KS	3.424%	(3.212%, 3.638%)	3.455%	0.000
MLE	4.207%	(3.932%, 4.507%)	3.900%	0.000
MLE Bias-Corrected	4.125%	(3.852%, 4.421%)	3.836%	0.000

6 Conclusion

The main considerations governing choice of a value-added estimator are efficiency, computational resource needs, and whether individual value-added scores or parameter estimates are desired.

More work is needed to understand which estimator is best for estimating individual effects, especially for a practitioner who cares only about ranking teachers and not about the magnitude of each teacher's score. It could be the case that a simple method works best: Previous studies have found that coefficients from fixed-effects regressions are very highly correlated with shrinkage value-added estimates (Kane *et al.*, 2013a). On the other hand, recent work suggests that machine learning methods perform well (Chalfin *et al.* (2016), Gramacy *et al.* (2016)). However, if a cardinal interpretation of value-added scores is desired, it becomes important to recover the right parameters in order to impose the proper degree of shrinkage.

For estimating the parameters of the distribution of value-added, the Kane and Staiger and modified Kane and Staiger estimators are the least computationally intensive; with N observations and K covariates, both are $O(NK^2)$. The most time-intensive step is running a least-squares regression. This algorithm then works with residuals, performing several quick $O(N)$ computations. The Kane and Staiger estimator comes with the most stringent identification requirements; it is only consistent when teachers are as good as randomly assigned. The modified-KS estimator is slower in practice since it requires using a within estimator, which makes sparse covariates dense.

Maximum likelihood is more statistically efficient, but less computationally efficient. Maximum likelihood estimates are lower-variance, and adding in the bias correction of Equation 8 greatly reduces bias. However, maximum likelihood estimation is significantly more time-intensive. Estimation iterates over variances (σ_μ^2, σ_θ^2, σ_ε^2) and coefficients (β, λ, α). Estimating $\hat{\beta}$ requires an $O(NK^2)$ regression using within-teacher variation *at every iteration*, and since variances have no closed-form solution, they must be numerically optimized..

Therefore, I recommend maximum likelihood estimation with a bias correction when it is computationally feasible, and the modified Kane and Staiger estimator under computational constraints.

References

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, **67** (2), 251–333.

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, **68** (2), 399–405.

Ballou, D., Sanders, W. and Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of educational and behavioral statistics*, **29** (1), 37–65.

Barnett, M. L., Olenski, A. R. and Jena, A. B. (2017). Opioid-Prescribing Patterns of

Emergency Physicians and Risk of Long-Term Use. *New England Journal of Medicine*, **376** (7), 663–673.

Briggs, D. and Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the. *National Education Policy Center*.

Buddin, R. (2011). Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J. and Mullainathan, S. (2016). Productivity and Selection of Human Capital with Machine Learning †. *American Economic Review*, **106** (5), 124–127.

Chamberlain, G. (1982). Panel Data. In *Handbook of Econometrics*, vol. II, Elsevier Science Publishers BV, pp. 1248–1313.

— (1984). Multivariate Regression Models for Panel Data. *Journal of Econometrics*, **18 (1982)**, 5–46.

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences*, **110** (43), 17176–17182.

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

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

Ellison, G. and Swanson, A. (2016). Do Schools Matter for High Math Achievement? Evidence from the American Mathematics Competitions. *American Economic Review*, **106** (6), 1244–1277.

Felch, J., Ferrell, S., Garvey, M., Lauder, T. S., Lauter, D., Marquis, J., Pesce, A., Poindexter, S., Schwencke, K., Shuster, B., Song, J. and Smith, D. (). Los Angeles Teacher Ratings. *Los Angeles Times*.

Feng, J. and Jaravel, X. (2016). Who Feeds the Trolls? Patent Trolls and the Patent Examination Process.

Gramacy, R. B., Taddy, M. and Tian, S. (2016). Hockey Player Performance via Regularized Logistic Regression. *arXiv preprint arXiv:1510.02172*.

Green, D. P. and Winik, D. (2010). Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism Among Drug Offenders*. *Criminology*, **48** (2), 357–387.

Guarino, C., Maxfield, M., Reckase, M., Thompson, P. and Wooldridge, J. (2014). An Evaluation of Empirical Bayes' Estimation of Value-Added Teacher Performance Measures.

Hanushek, E. A. and Rivkin, S. G. (2006). Chapter 18 Teacher Quality. In *Handbook of the Economics of Education*, vol. 2, Elsevier, pp. 1051–1078, dOI: 10.1016/S1574-0692(06)02018-6.

— and — (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, **100** (2), 267–271.

— and — (2012). The distribution of teacher quality and implications for policy. *Annu. Rev. Econ.*, **4** (1), 131–157.

Horowitz, J. L. (2001). The bootstrap. *Handbook of econometrics*, **5**, 3159–3228.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, **89** (5-6), 761–796.

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

Kane, T. J., McCaffrey, D. F., Miller, T. and Staiger, D. O. (2013a). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Seattle, WA: Bill and Melinda Gates Foundation*.

—, —, — and — (2013b). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Koedel, C. and Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Association for Education Finance and Policy*, **6** (1), 18–42.

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

Kozhimannil, K. B., Law, M. R. and Virnig, B. A. (2013). Cesarean Delivery Rates Vary Tenfold Among US Hospitals; Reducing Variation May Address Quality And Cost Issues. *Health Affairs*, **32** (3), 527–535.

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

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

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

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, **9** (1), 130–134.

Staiger, D. O. and Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, **24** (3), 97–118.

A Maximum Quasi-Likelihood Robustly Estimates Variances

If we assume the model of Section 3, in which data is drawn from some distribution $\mathcal{D}$ and we do not assume a functional form, then quasi-likelihood based on normality delivers consistent estimates of parameters $\eta = \left(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, \boldsymbol{\beta}, \boldsymbol{\lambda}, \alpha\right)$. Consider the normal model

$$\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j \sim N\left(\alpha + \boldsymbol{x}_j\boldsymbol{\beta} + \bar{\bar{x}}_j^T\boldsymbol{\lambda}, \Sigma(\eta, s_j)\right),$$

with the corresponding likelihood function $f\left(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \eta\right)$.

Lemma A.1.

$$\eta = \arg\max_\gamma \mathrm{E}_\mathcal{D} \log f\left(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \gamma\right) \tag{13}$$

Proof.

$$\begin{aligned}\mathrm{E}_\mathcal{D} \log f(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \gamma) = &-\frac{1}{2}\log\det\Sigma(\gamma, s_j) \\ &-\frac{1}{2}\mathrm{E}_\mathcal{D}\left[\left(\boldsymbol{y}_j - a - \boldsymbol{x}_j\boldsymbol{b} - \bar{\bar{x}}_j^T\boldsymbol{\ell}\right)^T \Sigma(\gamma, s_j)^{-1}\left(\boldsymbol{y}_j - a - \boldsymbol{x}_j\boldsymbol{b} - \bar{\bar{x}}_j^T\boldsymbol{\ell}\right) | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j\right]\end{aligned} \tag{14}$$

Since $\boldsymbol{\beta}$ corresponds to an unrestricted linear predictor, the values of α, $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ that maximize Equation 14 do not depend on Σ, so

$$\arg\max_{\boldsymbol{b},\boldsymbol{\ell}} \mathrm{E}_\mathcal{D} \log f(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j; \sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, \boldsymbol{b}, \boldsymbol{\ell}) = \boldsymbol{\beta}, \boldsymbol{\lambda}$$

After plugging in $\boldsymbol{b} = \boldsymbol{\beta}$ and $\boldsymbol{\ell} = \boldsymbol{\lambda}$, we can rewrite Equation 14 as

$$\begin{aligned}&\mathrm{E}_\mathcal{D} \log f(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j; \gamma, \boldsymbol{\beta}, \boldsymbol{\lambda}, \alpha) \\ &= -\frac{1}{2}\log\det\Sigma(\gamma, s_j) - \frac{1}{2}\mathrm{E}_\mathcal{D}\left[\left(\boldsymbol{y}_j - \alpha - \boldsymbol{x}_j\boldsymbol{\beta} - \bar{\bar{x}}_j^T\boldsymbol{\lambda}\right)^T \Sigma(\gamma, s_j)^{-1}\left(\boldsymbol{y}_j - \alpha - \boldsymbol{x}_j\boldsymbol{\beta} - \bar{\bar{x}}_j^T\boldsymbol{\lambda}\right) | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j\right] \\ &= -\frac{1}{2}\log\det\Sigma(\gamma, s_j) - \frac{1}{2}\mathrm{E}_\mathcal{D}\left[\mathrm{trace}\left(\boldsymbol{y}_j - \alpha - \boldsymbol{x}_j\boldsymbol{\beta} - \bar{\bar{x}}_j^T\boldsymbol{\lambda}\right)^T \Sigma(\gamma, s_j)^{-1}\left(\boldsymbol{y}_j - \alpha - \boldsymbol{x}_j\boldsymbol{\beta} - \bar{\bar{x}}_j^T\boldsymbol{\lambda}\right) | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j\right] \\ &= -\frac{1}{2}\log\det\Sigma(\gamma, s_j) - \frac{1}{2}\mathrm{trace}\left(\Sigma(\gamma, s_j)^{-1}\mathrm{E}_\mathcal{D}\left[\left(\boldsymbol{y}_j - \alpha - \boldsymbol{x}_j\boldsymbol{\beta} - \bar{\bar{x}}_j^T\boldsymbol{\lambda}\right)^T\left(\boldsymbol{y}_j - \alpha - \boldsymbol{x}_j\boldsymbol{\beta} - \bar{\bar{x}}_j^T\boldsymbol{\lambda}\right)\right]\right) \\ &= -\frac{1}{2}\log\det\Sigma(\gamma, s_j) - \frac{1}{2}\mathrm{trace}\left(\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)\right)\end{aligned}$$

Therefore,

$$\begin{aligned}\arg\max_\gamma \mathrm{E}_\mathcal{D} \log f(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j; \gamma, \boldsymbol{\beta}, \boldsymbol{\lambda}, \alpha) &= \arg\max_\gamma -\log\det\Sigma(\gamma, s_j) - \mathrm{trace}\left(\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)\right) \\ &= \arg\max_\gamma \log\det\left(\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)\right) - \mathrm{trace}\left(\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)\right)\end{aligned}$$

Let $\Sigma(\eta, s_j)^{1/2}$ be the symmetric, positive definite square root of the symmetric, positive definite matrix $\Sigma(\eta, \boldsymbol{x}_j)$, and let $\{e_i\}$ be the eigenvalues of $\Sigma(\eta, s_j)^{1/2}\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)^{1/2}$. Since $\Sigma(\eta, s_j)^{1/2}\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)^{1/2}$ is positive definite, all of its eigenvalues are positive.

$$\begin{aligned}
&\log\det\left(\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)\right) - \text{trace}\left(\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j\right) \\
&\quad = \log\det\left(\Sigma(\eta, s_j)^{1/2}\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)^{1/2}\right) - \text{trace}\left(\Sigma(\eta, s_j)^{1/2}\Sigma(\gamma, s_j)^{-1}\Sigma(\eta, s_j)^{1/2}\right) \\
&\quad = \log\Pi_i e_i - \sum_i e_i \\
&\quad = \sum_i \left(\log(e_i) - e_i\right) \qquad (15)
\end{aligned}$$

Equation 15 is maximized when all $e_i = 1$, which occurs when $\Sigma(\gamma, s_j) = \Sigma(\eta, s_j)$. As long as the teacher teaches multiple classes and at least one class has multiple students, the only value that solves this equation is $\gamma = \eta$.

□

B Closed-Form Likelihood and Intuitive Parameter Estimates

This section derives a closed-form solution for the likelihood. Subsection B.1 translates Equation 13, which is in terms of a determinant and an inverse of Σ, into an equation that contains integrals but no determinant or inverse. Subsection B.2 solves these integrals to give a tractable formula for the likelihood.

B.1 Matrices to Integrals

We can find a closed-form solution for the likelihood, without inverses, determinants, or integrals, by constructing a sum of independent variables that has the same distribution as $\boldsymbol{y}_j$. For each classroom c, define $\boldsymbol{\ell}_c$, a vector of ones with length equal to the number of students in classroom c, and for each teacher j number her classrooms $c = 1, 2, \ldots, C$. Define the following independent random variables:

$$\begin{aligned}
\mu_j &\sim N\left(\bar{\boldsymbol{x}}_j^T\boldsymbol{\lambda}, \sigma_\mu^2\right) \\
\theta_c &\sim N(0, \sigma_\theta^2) \\
\boldsymbol{\varepsilon}_j &\sim N\left(\alpha + \boldsymbol{x}_j^T\boldsymbol{\beta}, \boldsymbol{I}\sigma_\varepsilon^2\right) \qquad (16)
\end{aligned}$$

Stack the θ_c corresponding to each classroom into a vector Θ_j. The covariance of Θ_j is block diagonal, with diagonal blocks corresponding to each classroom:

$$\Theta_j = \begin{pmatrix} \boldsymbol{\theta}_1\ell_1 \\ \boldsymbol{\theta}_2\ell_2 \\ \vdots \\ \boldsymbol{\theta}_C\ell_C \end{pmatrix} \qquad \text{Var}(\Theta_j) = \sigma_\theta^2 B \qquad B = \begin{pmatrix} \ell_1\ell_1^T & 0 & 0 & 0 \\ 0 & \ell_2\ell_2^T & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \ell_C\ell_C^T \end{pmatrix}.$$

Summing the random variables of Equation 16 gives a new random variable that has the same distribution as $\boldsymbol{y}_j$:

$$\begin{aligned}
\mu_j + \Theta_j + \boldsymbol{\varepsilon}_j &\sim N\left(\alpha + \boldsymbol{x}_j^T\boldsymbol{\beta} + \bar{\bar{x}}_j^T\lambda, \boldsymbol{\ell}\boldsymbol{\ell}^T\sigma_\mu^2 + B\sigma_\theta^2 + \boldsymbol{I}\sigma_\varepsilon^2\right) \\
\boldsymbol{\ell}\boldsymbol{\ell}^T\sigma_\mu^2 + B\sigma_\theta^2 + \boldsymbol{I}\sigma_\varepsilon^2 &= \Sigma_j(\eta) \\
\mu_j + \Theta_j + \boldsymbol{\varepsilon}_j &\overset{D}{=} \boldsymbol{y}_j
\end{aligned}$$

Intuitively, μ_j affects all students taught by the same teacher, each θ_c affects each student in classroom c and is independent from all other θ_k, and each component of $\boldsymbol{\varepsilon}_j$ independently affects one student. Now we can use the distribution of $\mu_j + \Theta_j + \boldsymbol{\varepsilon}_j$ to come up with an alternative but equivalent description of the likelihood:

$$\begin{aligned}
f_{\boldsymbol{y}_j}(\boldsymbol{y}_j|\eta) &= f_{\mu_j+\Theta_j+\boldsymbol{\varepsilon}_j}(\boldsymbol{y}_j|\eta) \\
&= \int_\mu f(\mu) f_{\Theta_j+\boldsymbol{\varepsilon}_j}(\boldsymbol{y}_j - \mu)\,\mathrm{d}\mu \\
&= \int_\mu \phi\left(\mu; \sigma_\mu^2\right) f_{\Theta_j+\boldsymbol{\varepsilon}_j}(\boldsymbol{y}_j - \mu)\,\mathrm{d}\mu
\end{aligned} \tag{17}$$

Since the covariance matrix of $\Theta_j + \boldsymbol{\varepsilon}_j$ is block diagonal, we can write its probability density function as a product over the blocks, which correspond to classes:

$$f\left(\Theta_j + \boldsymbol{\varepsilon}_j\right) = \Pi_c f\left(\ell_c\theta_c + \varepsilon_c\right) \tag{18}$$

$$\begin{aligned}
\ell_c\theta_c + \varepsilon_c &\sim N\left(x_c^T\boldsymbol{\beta}, \ell_c\ell_c^T\sigma_\theta^2 + I\sigma_\varepsilon^2\right) \\
f_{\ell_c\theta_c+\varepsilon_c}(y_c - \mu) &= \int_\theta f(\theta) f_{\varepsilon_c}(y_c - \mu - \theta)\,\mathrm{d}\theta \\
&= \int_\theta \phi(\theta; \sigma_\theta^2)\Pi_{i\in I(c)}\phi\left(y_i - x_i^T\boldsymbol{\beta} - \mu - \theta; \sigma_\varepsilon^2\right)\mathrm{d}\theta
\end{aligned}$$

$$f_{\Theta_j+\boldsymbol{\varepsilon}_j}(\boldsymbol{y}_j - \mu) = \Pi_c \int_\theta \phi(\theta; \sigma_\theta^2)\Pi_{i\in I(c)}\phi\left(y_i - x_i^T\boldsymbol{\beta} - \mu - \theta; \sigma_\varepsilon^2\right)\mathrm{d}\theta \tag{19}$$

Plugging Equation 18 into Equation 17, we get a complete formula for the likelihood:

$$\begin{aligned}
&f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \theta, \boldsymbol{\beta}, \boldsymbol{\lambda}, \alpha\right) \\
&= \int_\mu \phi\left(\mu - \bar{\bar{x}}_j^T\boldsymbol{\lambda}; \sigma_\mu^2\right)\Pi_c\left(\int_\theta \phi\left(\theta; \sigma_\theta\right)\Pi_i\phi\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \mu - \theta - \alpha; \sigma_\varepsilon\right)\mathrm{d}\theta\right)\mathrm{d}\mu
\end{aligned} \tag{20}$$

B.2 Solving the Integrals

This section derives a closed-form solution for the likelihood using Equation 20 as a starting point. Two formulas will be used repeatedly. First, the product of the densities of n normal distributions with different means and the same variance is

$$\begin{aligned}\Pi_i \phi\left(\mu_i;\sigma\right) &\propto \sigma\phi\left(\tilde{\boldsymbol{\mu}}, \boldsymbol{I}\sigma^2\right)\phi\left(\bar{\mu},\sigma^2/n\right)\\ &= \sigma^{2-n}\phi\left(\sqrt{\sum_i \tilde{\mu}_i^2},\sigma^2\right)\phi\left(\bar{\mu},\sigma^2/n\right)\end{aligned} \tag{21}$$

where ϕ is the multivariate normal probability density function, and $\tilde{\mu}_i = \mu_i - \bar{\mu}$. The second useful identity is the product of the densities of two normal distributions, integrated over a translation of their means:

$$\int_\mu \phi(\mu - x_1;\sigma_1)\phi(\mu - x_2,\sigma_2)\,\mathrm{d}\mu = \phi\left(x_1 - x_2;\sqrt{\sigma_1^2+\sigma_2^2}\right) \tag{22}$$

The third useful formula is the product of n normal densities with different means and variances:

$$\begin{aligned}\Pi_c \phi(\mu_c;\sigma_c) &= \sqrt{\frac{1}{\sum_c 1/\sigma_c^2}}\phi\left(\tilde{\mu},\boldsymbol{I}\sigma^2\right)\phi\left(\bar{\mu},\frac{1}{\sum_c 1/\sigma_c^2}\right)\\ &\equiv \sqrt{\frac{1}{\sum_c h_c}}\phi\left(\tilde{\mu},\boldsymbol{I}(1/h)\right)\phi\left(\bar{\mu},\frac{1}{\sum_c h_c}\right)\end{aligned} \tag{23}$$

It is also helpful to set up notation for precision weights, which will pop out of derivations. The precision of the mean classroom error is

$$\frac{1}{h_c} \equiv \mathrm{Var}\left(\bar{y}_c - \bar{\boldsymbol{x}}_c\boldsymbol{\beta} - \mu_{j(c)}\right) = \sigma_\theta^2 + \sigma_\varepsilon^2/n_c. \tag{24}$$

Then use precision weights to create teacher-level means of y and x:

$$\bar{\bar{y}}_j \equiv \frac{\sum_{c:j(c)=j} h_c\bar{y}_c}{\sum_{c:j(c)=j} h_c} \qquad \bar{\bar{x}}_j \equiv \frac{\sum_{c:j(c)=j} h_c\bar{\boldsymbol{x}}_c}{\sum_{c:j(c)=j} h_c} \tag{25}$$

Letting n_c be the number of students in classroom c, Equation 21 implies

$$\Pi_i\phi(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \mu - \theta;\sigma_\varepsilon) = \sigma_\varepsilon^{2-n_c}\phi\left(\sqrt{\sum_i\left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}\right)^2};\sigma_\varepsilon^2\right)\phi\left(\bar{y}_c - \bar{x}_c^T\boldsymbol{\beta} - \mu - \theta - \alpha;\sigma_\varepsilon^2/n_c\right),$$

Adding in Equation 22,

$$\begin{aligned}&\int_\theta \phi\left(\theta;\sigma_\theta\right)\Pi_i\phi\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \mu - \theta - \alpha;\sigma_\varepsilon\right)\mathrm{d}\theta\\ &\qquad= \sigma_\varepsilon^{2-n_c}\phi\left(\sqrt{\sum_i\left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}\right)^2};\sigma_\varepsilon^2\right)\int_\theta\phi\left(\theta;\sigma_\theta\right)\phi\left(\bar{y}_c - \bar{x}_c^T\boldsymbol{\beta} - \mu - \theta - \alpha;\sigma_\varepsilon^2/n\right)\\ &\qquad= \sigma_\varepsilon^{2-n_c}\phi\left(\sqrt{\sum_i\left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}\right)^2};\sigma_\varepsilon^2\right)\phi\left(\bar{y}_c - \bar{x}_c^T\boldsymbol{\beta} - \mu - \alpha;\sigma_\vartheta^2 + \sigma_\varepsilon^2/n_c\right)\\ &\qquad= \sigma_\varepsilon^{2-n_c}\phi\left(\sqrt{\sum_i\left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}\right)^2};\sigma_\varepsilon^2\right)\phi\left(\bar{y}_c - \bar{x}_c^T\boldsymbol{\beta} - \mu - \alpha;1/h_c\right).\end{aligned}$$

From Equation 23,

$$\Pi_c \left(\bar{y}_c - \bar{x}_c^T \boldsymbol{\beta} - \mu - \alpha; 1/h_c\right) = \frac{1}{\sqrt{\sum_c h_c}} \phi\left(\tilde{\bar{y}}_j - \tilde{\bar{x}}_j^T \boldsymbol{\beta}, \boldsymbol{I}(1/h_j)\right) \phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T \boldsymbol{\beta} - \mu - \alpha, \frac{1}{\sum_c h_c}\right)$$

Therefore,

$$\begin{aligned}\Pi_c \int_\theta \phi\left(\theta; \sigma_\theta\right) \Pi_i \phi(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - \mu - \theta - \alpha; \sigma_\varepsilon) \, \mathrm{d}\theta =& \sigma_\varepsilon^{1+N_{\text{classes}}-N_{\text{students}}} \sqrt{\frac{\Pi_c h_c}{\sum_c h_c}} \exp\left(-\frac{1}{2} \sum_c \left(\tilde{\bar{y}} - \tilde{\bar{x}}^T \boldsymbol{\beta}\right)^2 h_c\right) \\ & \phi\left(\sqrt{\sum_i \left(\tilde{y}_i - \tilde{x}^T \boldsymbol{\beta}\right)^2}; \sigma_\varepsilon^2\right) \phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T \boldsymbol{\beta} - \mu - \alpha; \frac{1}{\sum_c h_c}\right)\end{aligned}$$

Applying Equation 22 again,

$$\int_\mu \phi\left(\mu - \bar{\bar{x}}_j^T \boldsymbol{\lambda}; \sigma_\mu^2\right) \phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T \boldsymbol{\beta} - \mu - \alpha, \frac{1}{\sum_c h_c}\right) \mathrm{d}\mu = \phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T (\boldsymbol{\beta} + \boldsymbol{\lambda}) - \alpha; \sigma_\mu^2 + \frac{1}{\sum_c h_c}\right)$$

The whole expression finally becomes

$$\begin{aligned}f\left(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \theta, \boldsymbol{\beta}, \boldsymbol{\lambda}, \alpha\right) =& \sigma_\varepsilon^{1+N_{\text{classes}}-N_{\text{students}}} \sqrt{\frac{\Pi_c h_c}{\sum_c h_c}} \exp\left(-\frac{1}{2} \sum_c \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T \boldsymbol{\beta}\right)^2 h_c\right) \\ & \phi\left(\sqrt{\sum_i \left(\tilde{y}_i - \tilde{x}_i^T \boldsymbol{\beta}\right)^2}; \sigma_\varepsilon^2\right) \phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T (\boldsymbol{\beta} + \boldsymbol{\lambda}) - \alpha; \sigma_\mu^2 + \frac{1}{\sum_c h_c}\right)\end{aligned}$$

The log-likelihood is

$$\begin{aligned}&\log f\left(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \theta, \boldsymbol{\beta}, \boldsymbol{\lambda}, \alpha\right) \\ &= (1 + N_{\text{classes}} - N_{\text{students}}) \log \sigma_\varepsilon + \frac{1}{2} \sum_c \log(h_c) - \frac{1}{2} \log\left(\sum_c h_c\right) - \frac{1}{2} \sum_c \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T \boldsymbol{\beta}\right)^2 h_c \\ &\quad + \log \phi\left(\sqrt{\sum_i \left(\tilde{y}_i - \tilde{x}_i^T \boldsymbol{\beta}\right)^2}, \sigma_\varepsilon^2\right) + \log \phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T (\boldsymbol{\beta} + \boldsymbol{\lambda}) - \alpha, \sigma_\mu^2 + \frac{1}{\sum_c h_c}\right) \\ &= (1 + N_{\text{classes}} - N_{\text{students}}) \log \sigma_\varepsilon + \frac{1}{2} \sum_c \log(h_c) - \frac{1}{2} \log\left(\sum_c h_c\right) - \frac{1}{2} \sum_c \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T \boldsymbol{\beta}\right)^2 h_c \\ &\quad - \log \sigma_\varepsilon - \frac{1}{2\sigma_\varepsilon^2} \sum_i \left(\tilde{y}_i - \tilde{x}_i^T \boldsymbol{\beta}\right)^2 - \frac{1}{2} \log\left(\sigma_\mu^2 + \frac{1}{\sum_c h_c}\right) - \frac{1}{2\left(\sigma_\mu^2 + \frac{1}{\sum_c h_c}\right)} \left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T (\boldsymbol{\beta} + \boldsymbol{\lambda}) - \alpha\right)^2 \\ &= (N_{\text{classes}} - N_{\text{students}}) \log \sigma_\varepsilon + \frac{1}{2} \sum_c \log(h_c) - \frac{1}{2} \log\left(\sum_c h_c\right) - \frac{1}{2} \sum_c \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T \boldsymbol{\beta}\right)^2 h_c \\ &\quad - \frac{1}{2\sigma_\varepsilon^2} \sum_i \left(\tilde{y}_i - \tilde{x}_i^T \boldsymbol{\beta}\right)^2 - \frac{1}{2} \log\left(\sigma_\mu^2 + \frac{1}{\sum_c h_c}\right) - \frac{1}{2\left(\sigma_\mu^2 + \frac{1}{\sum_c h_c}\right)} \left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T (\boldsymbol{\beta} + \boldsymbol{\lambda}) - \alpha\right)^2\end{aligned}$$

The log-likelihood for all teachers is

$$\begin{aligned}
&\sum_j \log f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \theta, \boldsymbol{\beta}, \boldsymbol{\lambda}, \alpha\right) \\
&\quad = \left(N_{\text{classes}} - N_{\text{students}}\right)\log\sigma_\varepsilon + \frac{1}{2}\sum_c \log h_c - \frac{1}{2}\sum_j \log\left(\sum_{c\in C(j)} h_c\right) - \frac{1}{2}\sum_c \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T\boldsymbol{\beta}\right)^2 h_c \\
&\qquad - \frac{1}{2\sigma_\varepsilon^2}\sum_i \left(\tilde{y}_i - \tilde{x}_i^T\boldsymbol{\beta}\right)^2 - \frac{1}{2}\sum_j \log\left(\sigma_\mu^2 + \frac{1}{\sum_{c\in C(j)} h_c}\right) - \sum_j \frac{1}{2\left(\sigma_\mu^2 + \frac{1}{\sum_{c\in C(j)} h_c}\right)}\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T\left(\boldsymbol{\beta}+\boldsymbol{\lambda}\right) - \alpha\right)^2.
\end{aligned}$$

If we wish to express the likelihood without integrating out teacher effects, we get

$$\begin{aligned}
f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, s_j; \eta\right) =& g\left(\eta\right) \\
&\int_\mu \phi\left(\mu - \bar{\bar{x}}_j^T\boldsymbol{\lambda}; \sigma_\mu^2\right)\phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T\boldsymbol{\beta} - \alpha - \mu; \frac{1}{\sum_c h_c}\right)\mathrm{d}\mu \qquad (26)
\end{aligned}$$

C Maximum Likelihood Bias Correction

Proof of Equation 7:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\hat{\boldsymbol{\lambda}}\right)\right] - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) &= \mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\hat{\boldsymbol{\lambda}}\right]\right) - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\
&= \mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\right]\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\
&= \mathbb{E}\left[\mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\bar{\bar{x}}\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\bar{\bar{x}}\right]\right) - \mathbb{E}\left[\bar{\bar{x}}\right]^T\mathrm{Cov}(\hat{\boldsymbol{\lambda}})\,\mathbb{E}\left[\bar{\bar{x}}\right] - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\
&= \mathbb{E}\left[\bar{\bar{x}}^T\,\mathrm{Cov}\left(\hat{\boldsymbol{\lambda}}\right)\bar{\bar{x}}\right] - \mathbb{E}\left[\bar{\bar{x}}\right]^T\mathrm{Cov}(\hat{\boldsymbol{\lambda}})\,\mathbb{E}\left[\bar{\bar{x}}\right] \\
&= \mathbb{E}\left[\left(\bar{\bar{x}} - \mathbb{E}\,\bar{\bar{x}}\right)^T\mathrm{Cov}\left(\hat{\boldsymbol{\lambda}}\right)\left(\bar{\bar{x}} - \mathbb{E}\,\bar{\bar{x}}\right)\right]. \qquad (27)
\end{aligned}$$

D Optimization,Gradient

This equation can easily be optimized numerically. The software package available at http://www.github.com/esantorella/tva iterates between estimating $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\lambda}}$, and $\hat{\alpha}$, which have closed-form solutions in terms of other parameters, and estimating σ_μ^2, σ_θ^2, and σ_ε^2 using L-BFGS.

D.1 Gradient

For compactness, let $\eta_j \equiv \frac{1}{\sum_{c\in C(j)} h_c}$.

$$
\begin{aligned}
\frac{\partial \mathrm{LL}}{\partial \sigma_\mu^2} =& \frac{1}{2}\sum_j \frac{1}{\left(\sigma_\mu^2+\eta_j\right)^2}\left(\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T(\boldsymbol{\beta}+\boldsymbol{\lambda})-\alpha\right)^2 - \left(\sigma_\mu^2+\eta_j\right)\right)\\
\frac{\partial \mathrm{LL}}{\partial \sigma_\theta^2} =& \frac{1}{2}\sum_c \frac{\partial h_c}{\partial \sigma_\theta^2}\left(\frac{1}{h_c} - \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T\boldsymbol{\beta}\right)^2\right)\\
&+\frac{1}{2}\sum_j\left(\sum_{c\in C(j)}\frac{\partial h_c}{\partial \sigma_\theta^2}\right)\left(-\frac{1}{1/\sigma_\mu^2+1/\eta_j} - \left(\frac{\eta_j}{\sigma_\mu^2+\eta_j}\right)^2\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T(\boldsymbol{\beta}+\boldsymbol{\lambda})-\alpha\right)^2\right)\\
\frac{\partial \mathrm{LL}}{\partial \sigma_\varepsilon^2} =& \frac{1}{2}\sum_c \frac{\partial h_c}{\partial \sigma_\varepsilon^2}\left(\frac{1}{h_c} - \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T\boldsymbol{\beta}\right)^2\right)\\
&+\frac{1}{2}\sum_j\left(\sum_{c\in C(j)}\frac{\partial h_c}{\partial \sigma_\varepsilon^2}\right)\left(-\frac{1}{1/\sigma_\mu^2+1/\eta_j} - \left(\frac{\eta_j}{\sigma_\mu^2+\eta_j}\right)^2\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T(\boldsymbol{\beta}+\boldsymbol{\lambda})-\alpha\right)^2\right)\\
&-\frac{1}{2}\frac{N_{\text{students}}-N_{\text{classes}}}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\varepsilon^4}\sum_i\left(\tilde{y}_i - \tilde{x}_i^T\boldsymbol{\beta}\right)^2\\
\frac{\partial \mathrm{LL}}{\partial \boldsymbol{\lambda}} =& \sum_j \frac{1}{\sigma_\mu^2+\eta_j}\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T(\boldsymbol{\beta}+\boldsymbol{\lambda})-\alpha\right)\bar{\bar{x}}_j\\
\frac{\partial \mathrm{LL}}{\partial \boldsymbol{\beta}} =& \sum_c\left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T\boldsymbol{\beta}\right)\tilde{\bar{x}}_c h_c + \frac{1}{\sigma_\varepsilon^2}\sum_i\left(\tilde{y}_i - \tilde{x}_i^T\boldsymbol{\beta}\right)\tilde{x}_i + \sum_j\frac{1}{\sigma_\mu^2+\eta_j}\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T(\boldsymbol{\beta}+\boldsymbol{\lambda})-\alpha\right)\bar{\bar{x}}_j\\
\frac{\partial \mathrm{LL}}{\partial \alpha} =& \sum_j \frac{1}{\sigma_\mu^2+\eta_j}\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T(\boldsymbol{\beta}+\boldsymbol{\lambda})-\alpha\right)
\end{aligned}
$$

E Bounding the Asymptotic Bias in the Kane and Staiger Procedure

We want to show that

$$
-b^T\,\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\right]b \le b\,\mathbb{E}\left[\bar{x}_{j1}\bar{x}_{j2}^T\right]b \le b^T\,\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\right]b \quad \forall b\in\mathcal{R}^k. \tag{28}
$$

To begin, note that since $\bar{x}_{j1}$ and $\bar{x}_{j2}$ are exchangeable, $\mathbb{E}\left[\bar{x}_{j1}\bar{x}_{j1}^T\right] - \mathbb{E}\left[\bar{x}_{j1}\bar{x}_{j2}^T\right]$ is positive semidefinite:

$$
\mathbb{E}\left[\bar{x}_{j1}\bar{x}_{j1}^T\right] - \mathbb{E}\left[\bar{x}_{j1}\bar{x}_{j2}^T\right] = \frac{1}{2}\mathbb{E}\left[\left(\bar{x}_{j1}-\bar{x}_{j2}\right)\left(\bar{x}_{j1}-\bar{x}_{j2}\right)^T\right].
$$

Also, $\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\right] - \mathbb{E}\left[\bar{x}_{j(i,c)1}\bar{x}_{j(i,c)1}^T\right]$ is positive semidefinite, since (letting $\boldsymbol{x}_i^1$ be a student drawn from classroom 1 of teacher $j(i,c)$):

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\right] - \mathbb{E}\left[\bar{x}_{j(i,c)1}\bar{x}_{j(i,c)1}^T\right] &= \mathbb{E}\left[\boldsymbol{x}_i^1\boldsymbol{x}_i^{1T}\right] - 2\,\mathbb{E}\left[\boldsymbol{x}_i^1\bar{x}_{j(i,c)1}\right] + \mathbb{E}\left[\bar{x}_{j(i,c)1}\bar{x}_{j(i,c)}^T\right]\\
&= \mathbb{E}\left[\left(\boldsymbol{x}_i^1 - \bar{x}_{j(i,c)1}\right)\left(\boldsymbol{x}_i^1 - \bar{x}_{j(i,c)1}\right)^T\right]
\end{aligned}
$$

Therefore, for any vector-valued b of the appropriate dimension,

$$b^T \left(\mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j1}^T \right] - \mathbb{E}\left[\bar{x}_{j1} \bar{x}_{j2}^T \right] \right) b \geq 0 \tag{29}$$

and

$$b^T \left(\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T \right] - E\left[\bar{x}_{j(i,c)1} \bar{x}_{j(i,c)1}^T \right] \right) b \geq 0. \tag{30}$$

Combining Equations 29 and 30,

$$b^T \mathbb{E}\left[\bar{x}_{j(i,c)1} \bar{x}_{j(i,c)2}^T \right] b \leq b^T \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T \right] \quad \forall b \in \mathcal{R}^k.$$

When this constraint binds, students are identical within a classroom, and classroom averages are the same for all classes taught by the same teacher. In the opposite case, when classrooms are perfectly anticorrelated, then

$$b^T \mathbb{E}\left[\bar{x}_{j(i,c)1} \bar{x}_{j(i,c)2}^T \right] b \geq -b^T \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T \right] b \quad \forall b \in \mathcal{R}^K.$$

F Figures

Figure 3: *Bias with Poisson-distributed outcomes: Results from 1000 Monte Carlo simulations. Error bars plot 95% confidence intervals based on the standard deviation of the bias.*

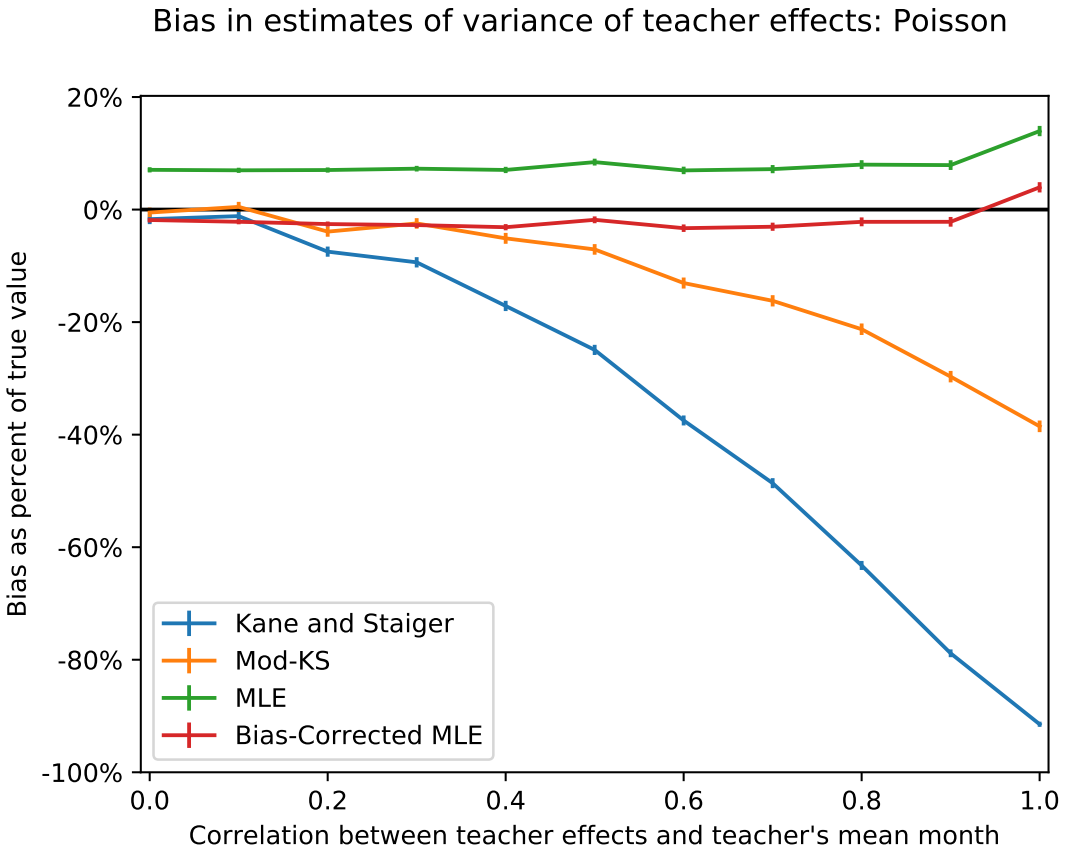


Figure 4: *Bias with t(3)-distributed outcomes: Results from 1000 Monte Carlo simulations. Error bars plot 95% confidence intervals based on the standard deviation of the bias.*