Measuring the Distribution of (Teacher) Value-Added

Elizabeth Santorella

January 11, 2018

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 that explicitly models the process of sorting of students to teachers. I use subsampling experiments based on a dataset of math teachers and math test scores in New York City to compare the bias and variance of each estimator, as well as the coverage properties of confidence intervals. Errors in variables can bias estimates, and the presence of covariates that do not vary within teacher leads to a lack of point identification; I discuss these issues and how they influence choice of estimator.

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 2% of the variance in student test scores. Value-added modeling is also used by school districts to rank teachers and make firing decisions. Although many papers have 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. As value-added estimation spreads to settings outside education, where data may be small and experimental validation infeasible, understanding identification and inference in value-added estimation is increasingly urgent.

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]

[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).

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 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 the distribution of individual effects. 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).

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, and derive asymptotic, analytic standard errors. I also develop a maximum (quasi-)likelihood estimator. I base empirical exercises off a dataset of teachers and students in New York City. Monte Carlo simulations based on this dataset, with simulated teacher effects and outcomes, confirm theoretical predictions about bias. In particular, these simulations show that bias varies with the correlation between teacher effects and covariates, and suggest that a bias-corrected maximum likelihood estimator nearly eliminates bias with no increase in variance. Next, by drawing small samples from the population of teachers and treating estimates from the whole data as the truth, as in Buchinsky (1995), I check the coverage probabilities of confidence intervals constructed using the asymptotic distributions of the estimators and find that they are slightly anti-conservative in small samples. Throughout, my focus is on the portion of variance that is due to variation in teacher quality. Although I derive formulas for individual value-added scores, I do not evaluate the accuracy of these scores. For clarity, I 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. In particular, subsection 3.4 notes that when some covariates do not vary within teacher, the variance of teacher effects is only partially identified without further

assumptions, and subsection 3.5 discusses how different estimators behave in the presence of errors in variables. Section 4 discusses the behavior of several procedures in Monte Carlo experiments, and Section 5 concludes with recommendations about which estimator to use.

1 Toy Model, Motivation

Why should policymakers care about the standard deviation of teacher effects? One plausible motivation is that the benefits of hiring, firing, and retraining teachers are increasing in the standard deviation of teacher effects. In particular, imagine that teacher value-added is accurately measured and is a sufficient statistic for teacher quality. Then the benefits of firing teachers in the bottom p fraction of teacher quality and replacing them with randomly-drawn teachers is a linear function of σ_μ.

Teacher quality is $\mu_j = \sigma_\mu \varepsilon_j$, where ε_j has mean zero and variance one. μ_j and ε_j have CDFs F_μ and F_ε. Then the benefit firing teachers in the bottom p fraction and replacing them with randomly-drawn teachers is

$$\begin{aligned}
\mathbb{E}\left[\mu_j\right] - \mathbb{E}\left[\mu_j | F(\mu_j) < p\right] &= 0 - \mathbb{E}\left[\mu_j | F_\mu(\mu_j) < p\right] \\
&= -\sigma_\mu \mathbb{E}\left[\varepsilon_j | F_\varepsilon(\varepsilon_j) < p\right] \\
&\propto \sigma_\mu.
\end{aligned}$$

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) [2]. However, later work has focused on "outcome-based" measures of teacher quality, treating quality as a 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

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

Table 1: *Estimates of the variance of teacher effects, $\widehat{\text{Var}}(\mu_j)$, 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	Reading
$\text{Var}\left(\mu_j\right)^{1/2}$	0.219	0.175
1 - forecast bias	0.905	1.089
	[0.552, 1.258]	[0.523, 1.655]

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 in educational settings (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. Consistent estimates of the variance of teacher effects, $\widehat{\text{Var}}(\mu_j)$, 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 $\widehat{\text{Var}}(\mu_j)$. The literature has typically interpreted forecast bias as a sign of insufficient controls for student-teacher sorting, but it can also reflect bias in $\widehat{\text{Var}}(\mu_j)$, 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. [3] 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 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

[3] In addition to the studies by Kane and Staiger (2008) and Chetty *et al.* (2014) cited below, Kane *et al.* (2013a) 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 $\text{Var}(\mu_j)$ was overestimated. They estimate $\text{Var}(\mu_j)$ to be 2.6% to 3.2% in math and 1% to 1.4% in reading.

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.

math scores by 0.219 standard deviations, with an analogous estimate of 0.175 for reading; their experimental results suggest that estimates are nearly unbiased. However, they are unable to rule out large degrees of bias. Estimates of about 0.1 standard deviations 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.* (2010) and found that a teacher one standard deviation above average improves math test scores by 0.27 standard deviations. That is, Buddin (2011) finds 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 and 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% from year to year.

In summary, two well-studied areas are whether the identification requirements of value-added models are obeyed and how accurately 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 the structural parameters that describe the distribution of parameter estimates.

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,

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
Sign of bias under baseline model	Up	?	Down	?
Closed-form solution	No	No	Yes	Yes
Closed-form asymptotic standard errors	MLE	?	GMM	GMM

and is otherwise negatively biased.

Observations are at the student level. Student i has classroom $c(i)$, teacher $j(i)$, test score y_i, and covariates $\boldsymbol{x}_i$ [4]. 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 $\bar{\bar{x}}_j \in \mathcal{R}^K$ [5]. $\boldsymbol{y}_j$ 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 c.

Test scores are generated according to

$$\boldsymbol{y}_j \equiv \mu_j + \boldsymbol{x}_j\boldsymbol{\beta} + \boldsymbol{\nu}_j. \tag{1}$$

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

In order to ascribe a casual interpretation to parameter estimates, we need sorting on observables. Sorting on observables requires the usual orthogonality restriction that $\boldsymbol{\nu}_j$ is orthogonal to covariates $\boldsymbol{x}_j$, so that we consistently estimate β. But sorting on observables requires not just orthogonality conditions, but *independence* conditions: unobservable shocks to test scores must be independent of assignments to teachers, so that $\boldsymbol{\nu}_j \perp\!\!\!\perp s_j | \boldsymbol{x}_j, \bar{\bar{x}}_j$. 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,

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

[5] $\bar{\bar{x}}_j$ is a precision-weighted mean, in a way that will be made clear.

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.

We can also model the relationship between teacher effects and covariates with teacher quality as a linear function of mean covariates:

$$\mu_j = \bar{\bar{x}}_j^T \boldsymbol{\lambda} + \tilde{\mu}_j, \quad \tilde{\mu}_j \perp \bar{\bar{x}}_j, s_j \tag{2}$$

$\boldsymbol{\lambda}$ 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.

$\text{Var}\left(\mu_j\right) = \text{Var}\left(\bar{\bar{x}}_j^T \boldsymbol{\lambda}\right) + \text{Var}(\tilde{\mu}_j)$ 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 $\text{Var}(\tilde{\mu}_j)$ is large, there are large differences in teacher quality that are not predictable from observables. When variance in $\bar{\bar{x}}_j \boldsymbol{\lambda}$ is large, there are large differences in teacher quality that are predictable by observables.

Combining Equations 1 and 2, $E_{\mathcal{D}}^* \left[\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j\right] = \boldsymbol{x}_j \boldsymbol{\beta} + \bar{\bar{x}}_j^T \boldsymbol{\lambda}$. When there are covariates that do not have within-teacher variation, $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are only partially identified. Section 3.4 derives bounds on $\text{Var}(\mu_j)$ and discusses how each estimator behaves in the presence of covariates that do not vary within teacher. In Sections 3.1 through 3.3, I assume that all covariates have within-teacher variation so that the model is point identified.

3.1 Maximum (Quasi-)Likelihood

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 with respect to s_j. Define $E\left[\boldsymbol{\nu}_j \boldsymbol{\nu}_j^T | s_j\right] \equiv \Sigma_j$. Denote parameters $\eta = \left(\boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2\right)$.

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

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. If the true values of parameters are η^* and maximum likelihood based on normality assumptions estimates η_F, then $\eta^* = \eta_F$. That is, consider the model

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

with the corresponding likelihood function $f\left(\boldsymbol{y}_j, \boldsymbol{x}_j, \bar{\tilde{x}}_j, s_j; \eta\right)$. Appendix A proves that $\eta_F = \arg\max_\eta \mathbb{E}_\mathcal{D} \log f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\tilde{x}}_j, s_j; \eta\right)$. Appendix B derives a relatively simple closed-form solution for the likelihood. A recurring theme is that important quantities are given in terms of classroom means $\bar{y}_c$, deviations from classroom means $\tilde{y}_i$, precisions $h_c = \frac{1}{\sigma_\theta^2 + \sigma_\varepsilon^2/|I(c)|}$, precision-weighted teacher-level means $\bar{\tilde{y}}_j$, and classroom-level deviations from teacher means $\tilde{\tilde{y}}_c$. Equations 21 through 25 define these terms.

Solving for the likelihood without integrating out teacher effects, as in Appendix Equation 29, 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 $\bar{\tilde{y}}_j \sim N\left(\mu_j + \bar{\tilde{x}}_j^T\boldsymbol{\beta}, \frac{1}{\sum_{c\in C(j)} h_c}\right)$, so the Empirical Bayes posterior of teacher j's value-added is

$$\mu_j \sim N\left(\frac{\sigma_\mu^2}{\sigma_\mu^2 + 1/\sum_{c\in C(j)} h_c}\overbrace{\left(\bar{\tilde{y}}_j - \bar{\tilde{x}}_j^T\boldsymbol{\beta}\right)}^{\text{residual}} + \frac{1/\sum_{c\in C(j)} h_c}{\sigma_\mu^2 + 1/\sum_{c\in C(j)} h_c}\underbrace{\bar{\tilde{x}}_j^T\boldsymbol{\lambda}}_{E[\mu_j|\bar{\tilde{x}}_j]}, \left(\frac{1}{\sigma_\mu^2} + \sum_{c\in C(j)} h_c\right)^{-1}\right) \tag{3}$$

The solutions for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}$ 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:

$$\hat{\boldsymbol{\beta}} = \arg\min_b \frac{1}{\hat{\sigma}_\varepsilon^2}\sum_j \sum_{c\in C(j)} \sum_{i\in I(c)} \left(\tilde{y}_i - \tilde{x}_i^T\boldsymbol{b}\right)^2 + \sum_j \sum_{c\in C(j)} \hat{h}_c\left(\bar{\tilde{y}}_c - \bar{\tilde{x}}_c^T\boldsymbol{b}\right)^2 \tag{4}$$

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

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\ell}} \sum_j \left(\frac{1}{\sum_c \hat{h}_c} + \hat{\sigma}_\mu^2\right)^{-1}\left(\bar{\tilde{y}}_j - \bar{\tilde{x}}_j^T\hat{\boldsymbol{\beta}} - \bar{\tilde{x}}_j^T\boldsymbol{\ell}\right)^2 \tag{5}$$

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 4 collapses to

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

and Equation 5 becomes

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

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

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

3.1.1 Inference

Since this is quasi-likelihood, asymptotic inference is (conceptually) easy. The standard errors in this paper use the robust "sandwich" standard errors, as well as reporting non-robust standard errors based on the inverse of the Fisher information matrix. Appendix D gives the first derivative of the likelihood function. Standard errors in this paper were calculated using an analytic Fisher information matrix and numerical second derivative.

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:

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

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 30, the sample analog is biased upwards by

$$\mathbb{E}\left[\text{Var}\left(\bar{\bar{x}}_j^T \hat{\boldsymbol{\lambda}} | \hat{\boldsymbol{\lambda}} \right) \right] - \text{Var}\left(\bar{\bar{x}}^T \boldsymbol{\lambda} \right) = \mathbb{E}\left[\left(\bar{\bar{x}}_j - \mathbb{E}\, \bar{\bar{x}}_j \right)^T \text{Cov}\left(\hat{\boldsymbol{\lambda}} \right) \left(\bar{\bar{x}}_j - \mathbb{E}\, \bar{\bar{x}}_j \right) \right] . \tag{6}$$

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

$$\widehat{\text{Var}}\left(\mu_j \right) = \overbrace{\frac{1}{n} \sum_j \left(\bar{\bar{x}}_j^T \hat{\boldsymbol{\lambda}} \right)^2 - \left(\frac{1}{n} \sum_j \bar{\bar{x}}_j^T \hat{\boldsymbol{\lambda}} \right)^2}^{\text{predictable variance}} + \underbrace{\hat{\sigma}_\mu^2}_{\text{unpredictable variance}} - \overbrace{\frac{1}{n} \sum_j \left(\bar{\bar{x}}_j^T - \frac{1}{J} \sum_k \bar{\bar{x}}_k \right)^T \hat{\Sigma}_\lambda \left(\bar{\bar{x}}_j - \frac{1}{J} \sum_k \bar{\bar{x}}_k \right)}^{\text{bias correction}} . \tag{7}$$

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. Their main specification not reject the hypothesis that the scores were forecast unbiased, but they lacked power to rule out a moderate degree of bias, and other specifications suggested that value-added scores could 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 replaces the random teacher effects assumption with a fixed effects or correlated random effects assumption, and consistently estimates the model laid out in Section 3.1 under only a sorting on observables requirement.

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 estimate of $\boldsymbol{\beta}$ if and only if teacher effects are uncorrelated with covariates; otherwise, this estimate will suffer from omitted variable bias:

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

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:

$$\hat{\sigma}^2_\mu = \frac{2}{\sum_j |C(j)| \left(|C(j)| - 1\right)} \sum_j \sum_{c,c' \in C(j): c \neq c'} \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}}\right) \left(\bar{y}_{c'} - \bar{\boldsymbol{x}}_{c'}^T \hat{\boldsymbol{\beta}}\right) \tag{8}$$

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:

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

$\hat{\sigma}_\theta^2$ 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 $\eta = (\boldsymbol{\beta}, \sigma_\mu^2, \sigma_\varepsilon^2, \sigma_\theta^2)$. The moment function, which is at the teacher level, is

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

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

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

3.2.3 Individual Teacher Effects

Although a teacher's mean residuals are 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 to generate a mean squared error-minimizing estimate of μ_j:

$$\hat{\mu}_j = \hat{\rho}_j \frac{\sum_{c \in C(j)} h_c \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}} \right)}{\sum_{c \in C(j)} h_c}$$

$$\hat{\rho}_j = \arg\min_{\rho} \mathbb{E} \left[\left(\rho \frac{\sum_{c:j(c)=j} \hat{h}_c \left(\bar{y}_c - \bar{\boldsymbol{x}}_c^T \hat{\boldsymbol{\beta}} \right)}{\sum_{c \in C(j)} h_c} - \mu_j \right)^2 \right] = \frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2 + \frac{1}{\sum_{c \in C(j)} \hat{h}_c}} \tag{9}$$

Kane and Staiger note that when μ, θ, and ε are normally distributed, Equation 9 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 9 equals Equation 3, 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 less weight on students in larger classrooms, due to the presence of classroom-level shocks.

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; here I demonstrate that omitted variable bias in $\hat{\boldsymbol{\beta}}$ leads to an asymptotic *negative* bias in $\widehat{\text{Var}}(\mu_j)$. Intuitively, variation in teacher effects that is correlated with student characteristics is incorrectly attributed to the student characteristics. Equation 8 gives

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

Appendix E shows that in the special case where each teacher teaches the same number of classrooms and each classroom has the same number of students, the bias can be bounded:

$$-3 \leq \frac{\text{Bias}(\widehat{\text{Var}}(\mu_j))}{\frac{\sum_j \bar{\tilde{x}}_j^T \mu_j}{N_{\text{teachers}}} \left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}^{-1} \right) \frac{\sum_j \bar{\tilde{x}}_j \mu_j}{N_{\text{teachers}}}} \leq -1 \tag{10}$$

Equation 10 makes several facts apparent. The estimator is always negatively biased even asymptotically, and bias is more severe when sorting is strong. This happens when $\boldsymbol{\lambda}$ is large in magnitude.

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 [6]. 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 (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). Guarino *et al.* (2014) explain that "estimators that include the teacher assignment indicators along with the covariates in a multiple regression analysis" perform better. Using within-teacher variation means that $\hat{\beta} \to_p \beta$, which in turn implies that $\widehat{\text{Var}}(\mu_j) \to_p \text{Var}(\mu_j)$. However, Section 3.5 shows that in the presence of errors in variables, when it is not possible to perfectly account for sorting, both estimators' estimates of $\hat{\beta}$ will be affected by attenuation bias, but the bias will be more severe for the Modified-KS estimator.

3.3.1 Inference

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

$$y_i^{\text{within}} \equiv y_i - \frac{1}{\sum_{c \in C(j)} |I(c)|} \sum_{c \in C(j)} \sum_{i' \in I(c')} y_{i'}$$

$$x_i^{\text{within}} \equiv x_i - \frac{1}{\sum_{c \in C(j)} |I(c)|} \sum_{c \in C(j)} \sum_{i' \in I(c')} x_{i'}$$

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

3.4 Covariates without within-teacher variation

Analysis up to this point has assumed that it is possible to separately identify $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. , This model is not point identified when a column of x_j is constant within each teacher. In

[6] 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.

that case, adding a constant to the corresponding component of β and subtracting it from the corresponding component of λ would yield the same prediction of $\boldsymbol{y}_j$. For example, a constant term could not be identified in this model, as it is impossible to distinguish the level of teacher effects from the level of the other factors influencing student achievement. The difficulty caused by a constant can be easily sidestepped by centering x_i and y_i around zero, but other covariates that don't vary within teacher cause more meaningful problems. For example, if all teachers have a constant gender throughout the sample, it is not possible to distinguish a world in which female teachers are better from one in which female teachers are assigned better students. Imagine that all students obtain exactly the same test scores. In such a case it would be tempting to claim that $\text{Var}(\mu_j) = 0$, but it could be the case that female teachers are much better than male teachers, and female teachers are assigned students who are unobservably more difficult, with negative values of ν_i. [7]

To see the effects of introducing covariates that do not vary within teacher, recall that the original model is given by

$$\begin{aligned} y_j &= \mu_j + x_j\beta + \nu_j, & \nu_j &\perp 1, x_j \\ \mu_j &= \bar{\bar{x}}_j^T\lambda + \tilde{\mu}_j, & \tilde{\mu}_j &\perp 1, \bar{\bar{x}}_j \\ \nu_j \underline{\perp\!\!\!\perp} s_j | x_j, \bar{\bar{x}}_j & \text{sorting on observables} \end{aligned}$$

The predicted value of y_j is

$$E^*\left[y_j | x_j, \bar{\bar{x}}_j, s_j\right] = x_j\boldsymbol{\beta} + \bar{\bar{x}}_j^T\boldsymbol{\lambda}$$

If there are K_1 covariates that do not vary within teacher, the model is only identified after fixing K_1 scalars. To see this, let $\ell \in \mathcal{R}^{n_s}$ be a vector of ones, and write $\boldsymbol{x}_j$ as $\boldsymbol{x}_j = (\boldsymbol{x}_j^0, x_j^1\boldsymbol{\ell}, \ldots, x_j^{K_1}\boldsymbol{\ell})$, where each column of $\boldsymbol{x}_j^0 \in \mathcal{R}^{n_s \times (K-K_1)}$ has within-teacher variation. Then the following equations are equally consistent with the data for all $\boldsymbol{\alpha} \in \mathcal{R}^{K_1}$:

$$\begin{aligned} E^*_{\mathcal{D}}\left[\boldsymbol{y}_j | \mu_j, \boldsymbol{x}_j^0, x_j^1, \ldots, x_j^{K_1}, s_j\right] &= \mu_j + \boldsymbol{x}_j^0\tilde{\beta} + \sum_{k=1}^{K_1}\alpha_k x_j^k\gamma_k \\ &\equiv \mu_j + \boldsymbol{x}_j\beta, \quad \boldsymbol{\beta}^T = (\tilde{\beta}^T, \alpha_1\gamma_1, \ldots, \alpha_{K_1}\gamma_{K_1}) \\ E^*_{\mathcal{D}}\left[\mu_j | \bar{\bar{x}}_j^0, x_j^1, \ldots, x_j^{K_1}, s_j\right] &= \bar{\bar{x}}_j^{0T}\tilde{\lambda} + \sum_{k=1}^{K_1}(1-\alpha_k)x_j^k\gamma_k \\ &\equiv \bar{\bar{x}}_j^T\boldsymbol{\lambda}, \quad \boldsymbol{\lambda}^T = (\tilde{\lambda}^T, (1-\alpha_1)\gamma_1, \ldots, (1-\alpha_{K_1})\gamma_{K_1}) \end{aligned} \tag{11}$$

$\tilde{\beta}$, $\tilde{\lambda}$, and $\boldsymbol{\gamma}$ are all point identified, but $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are not without knowing $\boldsymbol{\alpha}$. $\text{Var}\left(\mu_j\right)$, the variance in teacher quality, depends on α:

[7] Randomly assigning students to teachers would identify that either β^k or λ^k is zero for many attributes x^k by enforcing that there is no relationship between teacher quality and student attributes or between student quality and teacher attributes. For example, the component of λ corresponding to students' previous test scores would be zero: there would be no association between teacher quality and student attributes. And the component of β corresponding to teacher gender would be zero: There would be no relationship between teacher gender and student quality.

$$\begin{aligned}\text{Var}\left(\mu_j\right) &= \text{Var}\left(\bar{\tilde{x}}_j^T \boldsymbol{\lambda}\right) + \sigma_\mu^2 \\ &= \text{Var}\left(\bar{\tilde{x}}_j^{0T} \tilde{\lambda} + \sum_{k=1}^{K_1} (1-\alpha_k) x_j^k \gamma_k\right) + \sigma_\mu^2\end{aligned}$$

How can the previously described estimators accommodate uncertainty about $\boldsymbol{\alpha}$? In the likelihood and modified-KS estimators, we can find a particular solution by setting $\boldsymbol{\alpha} = (0, \ldots, 0)^T$, and then find a solution for any other value of $\boldsymbol{\alpha}$. Since the Kane and Staiger estimator is only consistent when $\boldsymbol{\lambda} = (0, \ldots, 0)^T$, it is nonsensical to ask how it can handle uncertainty about $\boldsymbol{\alpha}$.

Among the remaining estimators, quasi-ML and modified-KS, quasi-ML is better able to explore the implications of varying the sorting parameter $\boldsymbol{\alpha}$ because it fully models the sorting process. Quasi-ML can generate an estimate of $\widehat{\text{Var}}(\mu_j)$ for any value of $\boldsymbol{\alpha}$, while mod-KS can only give estimates for $\boldsymbol{\alpha} = (0, \ldots, 0)$ or the value that minimizes $\widehat{\text{Var}}(\mu_j)$. Asymptotically, both quasi-ML and modified-KS estimate the same lower bound of the identified set (which is not bounded above).

3.4.1 Maximum (Quasi-)Likelihood

Using the likelihood model, we can obtain maximum likelihood estimates for $\hat{\beta}$ and $\hat{\gamma}$ for a particular value of $\boldsymbol{\alpha}$, back out the point-identified parameters $\tilde{\beta}$, $\tilde{\lambda}$, and $\tilde{\gamma}$, and then find $\widehat{\text{Var}}(\mu_j)$ for a variety of values of $\boldsymbol{\alpha}$.

First, set the last K_1 components of $\hat{\beta}$ to zero, setting $\boldsymbol{\alpha} = (0, \ldots, 0)$, by attributing variation due to $(x_j^1, \ldots, x_j^{K_1})$ to teacher quality. Since $\hat{\beta}$ is identified off within-teacher variation, this corresponds to saying coefficients on variables with no within-teacher variation are zero. Now we can transform these estimands to those corresponding to any other value of α: Set $\hat{\tilde{\beta}}$ to the first $K - K_1$ components of $\hat{\beta}$, $\hat{\tilde{\lambda}}$ to the first $K - K_1$ components of $\hat{\lambda}$, and $\hat{\gamma}$ to the last K_1 components of $\hat{\lambda}$.

Then we can find $\widehat{\text{Var}}(\mu_j)(\boldsymbol{\alpha})$ for any $\boldsymbol{\alpha}$ as

$$\widehat{\text{Var}}(\mu_j) = \hat{\sigma}_\mu^2 + \widehat{\text{Var}}\left(\bar{\tilde{x}}_j^{0T} \tilde{\lambda} + \sum_{k=1}^{K_1} (1-\alpha_k) x_j^k \gamma_k\right).$$

By varying α, we can see that there are various values of $\text{Var}(\mu_j)$ that are consistent with the point-identified parameters $\tilde{\lambda}$ and $\boldsymbol{\gamma}$.

$$
\begin{aligned}
\widehat{\text{Var}}\left(\mu_j\right)^{\text{MLE}}_{\boldsymbol{\alpha}=(0,\ldots,0)} &= \hat{\sigma}^2_\mu + \widehat{\text{Var}}\left(\bar{\tilde{x}}^{0T}_j\tilde{\lambda} + x^{1T}_j\boldsymbol{\gamma}\right) \\
\widehat{\text{Var}}\left(\mu_j\right)^{\text{MLE}}_{\boldsymbol{\alpha}=(1,\ldots,1)} &= \hat{\sigma}^2_\mu + \widehat{\text{Var}}\left(\bar{\tilde{x}}^{0T}_j\tilde{\lambda}\right) \\
\min_{\alpha}\widehat{\text{Var}}\left(\mu_j\right)^{\text{MLE}}_{\boldsymbol{\alpha}} &= \sigma^2_\mu + \text{Var}\left(\bar{\tilde{x}}^{0T}_j\tilde{\lambda} - x^{1T}_j\left(\sum_{j'}x^1_{j'}x^{1T}_{j'}\right)^{-1}\left(\sum_{j'}x^1_{j'}\bar{\tilde{x}}^{0T}_{j'}\tilde{\lambda}\right)\right) \\
&\rightarrow_p \sigma^2_\mu + \text{Var}\left(\bar{\tilde{x}}^{0T}_j\tilde{\lambda} - E^*\left[\bar{\tilde{x}}^{0T}_j\tilde{\lambda}|x^1_j,\ldots,x^{K_1}_j\right]\right) \\
\max_{\alpha}\widehat{\text{Var}}\left(\mu_j\right)^{\text{MLE}} &= \infty
\end{aligned}
$$

First, in the case that $\boldsymbol{\alpha}$ consists of all zeros, any variation that could potentially be attributed to either teachers (λ) or students (β) is attributed to teachers. In the case that $\boldsymbol{\alpha}$ consists of all ones, this variation is attributed to students. The lowest estimate of the variance of teacher effects comes from choosing each component of $\boldsymbol{\alpha}$ so that $\sum_{k=1}^{K_1}(1-\alpha_k)x^k_j\gamma_k$ cancels out as much of the variation in $\bar{\tilde{x}}^{0T}_j\tilde{\lambda}$ as possible. And as $\boldsymbol{\alpha}$ tends towards infinity or negative infinity, the variance of teacher effects approaches infinity. This is similar to a situation in which female teachers are vastly better than male teachers but are assigned students with vastly lower values of ν_i. Restricting each component of $\boldsymbol{\alpha}$ to lie in $[0,1]$ allows for only "positive" sorting, in which factors that are positively correlated with student achievement are also positively correlated with teacher quality.

3.4.2 Modified-KS

With the modified-KS estimator, like the quasi-ML estimator, we can find a particular solution for $\hat{\beta}$ by setting coefficients on covariates without within-teacher variation to 0. This corresponds to assuming that each component of α is zero. Then $\hat{\tilde{\beta}}$ equals the first $K - K_1$ components of $\tilde{\beta}$. However, since the modified-KS estimator does not explicitly model sorting, γ is unknown. Therefore, we cannot find $\widehat{\text{Var}}(\mu_j)$ as a function of $\boldsymbol{\alpha}$, but rather as a function of $\boldsymbol{\phi} \equiv (\alpha_1\gamma_1,\ldots,\alpha_{K_1}\gamma_{K_1})^T$, which is less easy to interpret.

$$
\widehat{\text{Var}}(\mu_j)^{\text{Mod}-\text{KS}}_{\boldsymbol{\phi}} = \frac{1}{\sum_j |C(j)|\left(|C(j)|-1\right)}\sum_j\sum_{c\in C(j)}\sum_{c'\in C(j),c'\neq c}\left(\bar{y}_c - \bar{x}^{0T}_c\hat{\tilde{\beta}} - \boldsymbol{x}^{1T}_j\boldsymbol{\phi}\right)\left(\bar{y}_{c'} - \bar{x}^{0T}_{c'}\hat{\tilde{\beta}} - \boldsymbol{x}^{1T}_j\boldsymbol{\phi}\right)
$$

Although we cannot estimate $\widehat{\text{Var}}(\mu_j)$ as a function of $\boldsymbol{\alpha}$ due to the presence of $\boldsymbol{\gamma}$, we can still plug in $\boldsymbol{\alpha} = (0,\ldots,0)^T$ and take the limit as at least one component of $\boldsymbol{\alpha}$ goes to infinity:

$$\begin{aligned}
\widehat{\text{Var}}\left(\mu_j\right)^{\text{mod}-\text{KS}}_{\phi=(0,\ldots,0)} &= \frac{1}{\sum_j |C(j)|\left(|C(j)|-1\right)} \sum_j \sum_{c,c' \in C(j), c \neq c'} \left(\bar{y}_c - \bar{x}_c^{0T}\hat{\tilde{\beta}}\right)\left(\bar{y}_{c'} - \bar{x}_{c'}^{0T}\hat{\tilde{\beta}}\right) \\
&\to_p \text{plim}\left(\widehat{\text{Var}}(\mu_j)^{\text{MLE}}_{\alpha=(0,\ldots,0)^T}\right) \\
\max_{\phi} \widehat{\text{Var}}\left(\mu_j\right)^{\text{mod}-\text{KS}} &= \infty
\end{aligned}$$

The minimum value of $\widehat{\text{Var}}(\mu_j)$ is given by setting each ϕ to minimize the sample covariance between residuals in different classes:

$$\arg\min_{\phi} \widehat{\text{Var}}(\mu_j)_{\phi} = \left(\sum_j \sum_{c \in C(j)} \sum_{c' \in C(j), c' \neq c} x_j^0 x_j^{0T}\right)^{-1} \left(\sum_j \sum_{c \in C(j)} \sum_{c' \in C(j), c' \neq c} x_j^0 \left(\bar{y}_c - \bar{x}_c^T \hat{\tilde{\beta}}\right)\right)$$

So asymptotically, modified-KS estimates the same lower bound as quasi-ML, despite not fully modeling the sorting process:

$$\begin{aligned}
\min_{\phi} \widehat{\text{Var}}(\mu_j) &\to_p \text{Cov}\left(\bar{y}_c - \bar{x}_c^{0T}\tilde{\beta} - E^*\left[\bar{y}_c - \bar{x}_c^{0T}\tilde{\beta} | \boldsymbol{x}_j^1\right], \bar{y}_{c'} - \bar{x}_{c'}^{0T}\tilde{\beta} - E^*\left[\bar{y}_{c'} - \bar{x}_{c'}^{0T}\tilde{\beta} | \boldsymbol{x}_j^1\right] \middle| c, c' \in C(j), c \neq c'\right) \\
&= \text{plim}\left(\min_{\boldsymbol{\alpha}} \widehat{\text{Var}}(\mu_j)^{\text{MLE}}_{\boldsymbol{\alpha}}\right)
\end{aligned}$$

3.5 Errors in variables

In educational settings, it is likely that teachers are sorted to students on the basis of unobservable characteristics like ability or parent involvement that are only roughly captured by controls. To investigate how errors in variables affects estimates, this section studies the implications of sorting on student ability that is only approximately captured by test scores.

Posit that ability is a mean-zero scalar. $\boldsymbol{y}_j \in \mathcal{R}^{n_s}$, $\boldsymbol{z}_j \in \mathcal{R}^{n_s}$, $\boldsymbol{x}_j \in \mathcal{R}^{n_s}$, and β, λ, $\bar{\tilde{x}}_j$, and $\bar{\tilde{z}}_j$ are scalars. The data is generated by

$$\begin{aligned}
\boldsymbol{y}_j &= \mu_j + \boldsymbol{z}_j\beta + \boldsymbol{\nu}_j \\
\mu_j &= \bar{\tilde{z}}_j\lambda + \tilde{\mu}_j
\end{aligned}$$

with the usual orthogonality restrictions. We do not observe $\boldsymbol{z}_j$, only $\boldsymbol{x}_j$. Each component x_i of $\boldsymbol{x}_j$ is generated with independent measurement error: $x_i = z_i + \sigma\varepsilon_i$, where ε_i has mean zero and variance one. If we control for $\boldsymbol{x}_j$ instead of $\boldsymbol{z}_j$, how does this affect our estimates of $\text{Var}(\mu_j)$?

The Kane and Staiger estimator estimates

$$\begin{aligned}
\hat{\beta}^{\text{KS}} &\to_p \frac{\text{Cov}(y_i, x_i)}{\text{Var}(x_i)} \\
&= \frac{\text{Cov}(\bar{\tilde{z}}_{j(i)}, z_i)\lambda + \text{Var}(z_i)\beta}{\text{Var}(x_i)} \\
&= \frac{\text{Var}(z_i)}{\text{Var}(x_i)}\beta + \frac{\text{Var}(\bar{\tilde{z}}_j)}{\text{Var}(\bar{\tilde{x}}_j)}\lambda
\end{aligned}$$

When $\hat{\beta}$ and $\hat{\lambda}$ have the same sign, as seems likely in an educational context, then without errors in variables $\hat{\beta}$ will be biased away from zero. Sorting on observables can push $\hat{\beta}$ towards zero, a case of two wrongs making a right. Without errors in variables, $\text{Var}(\mu_j)$ was biased downwards; now its bias cannot be signed.

The modified-KS estimator uses within-teacher variation:

$$\hat{\beta} \rightarrow_p \frac{\text{Var}(z_i^{\text{demeaned}})}{\text{Var}(x_i^{\text{demeaned}})}\beta$$

3.5.1 Calibration

We can calibrate σ and other parameters needed to assess the impact of errors in variables using data from New York City, described in more detail below. The test-retest reliability of standardized tests is approximately 0.8. When the variance of test scores x_i is normalized to 1, this implies that $\sigma^2 = .2$. Letting "classrooms" c be teacher-year-grade units, we can also back out that latent ability is the sum of a teacher-level component with variance 0.34, a teacher-year-grade component with variance 0.04, and an individual-level component with variance 0.42.

In other words, using notation consistent with that previously used to describe teacher-level, and classroom-level variables,

$$\begin{aligned}
z_i &= \bar{\bar{z}}_{j(i)} + \tilde{\bar{z}}_{c(i)} + \tilde{z}_i \\
\text{Var}(\bar{\bar{z}}_j) &= 0.34 \\
\text{Var}(\tilde{\bar{z}}_c) &= 0.04 \\
\text{Var}(\tilde{z}_i) &= 0.42 \\
x_i &= z_i + \sigma\varepsilon_i \\
\sigma^2 &= 0.2
\end{aligned}$$

Therefore, in the Kane and Staiger estimator, assuming $\lambda = 0$,

$$\begin{aligned}
\lim_{N_{\text{teachers}}\rightarrow\infty, N_{\text{students per teacher}}\rightarrow\infty} \hat{\beta}^{\text{KS}} &= 0.8\beta \\
\lim_{N_{\text{teachers}}\rightarrow\infty, N_{\text{students per teacher}}\rightarrow\infty} \hat{\beta}^{\text{mod-KS}} &= \frac{\text{Var}(z_i) - \text{Var}(\bar{\bar{z}}_j)}{\text{Var}(x_i) - \text{Var}(\bar{\bar{z}}_j)}\beta \\
&= 0.70\beta
\end{aligned}$$

Since students with the same teacher tend to be similar in ability, using the modified Kane and Staiger estimator instead of the Kane and Staiger estimator substantially worsens attenuation bias.

4 Monte Carlo Experiments

4.1 Data and real results

In this section, I use a dataset of math teachers in New York City to estimate the variance of math teacher effects on math test scores. The data runs from 2006 to 2015 and includes over 10 million student-test score observations. In accordance with previous teacher value-added literature, I control for demographic information and student achievement: gender, whether the student is disabled, whether the student is in a specialized class for English language learners, free or reduced price lunch status, lagged attendance, and lagged test scores. In order to always control for lagged test scores, third graders and students who recently moved to the district are not included. Summary statistics for the whole data are given in Table 3.

Table 3: *Summary statistics, New York City public schools.*

	Mean	St. Dev	Min	Max	Missing
Grade	5.65	3.94	-1	12	0.07%
Year	2009.39	3.11	2005	2015	0%
Disabled	0.15	0.36	0	1	0%
Female	0.49	0.50	0	1	0%
English Language Learner	0.13	0.34	0	1	0%
Free Lunch	0.77	0.42	0	1	0%
Days absent	16.19	22.16	0	187	6.39%
Days present	154.88	35.20	0	329	6.39%
Days Absent Lag (Z-Score)	0.03	0.97	-16.45	1.40	28.17%
Math Score (Z-Score)	0.01	1.00	-10.00	3.96	60.38%
Math score lag (Z-Score)	0.00	0.99	-10.00	3.96	66.3%
ELA Score (Z-Score)	0.00	1.00	-11.10	7.76	61.17%
ELA Score Lag (Z-Score)	-0.00	0.99	-11.10	7.76	67.14%
4-Year Graduation	0.59	0.49	0	1	70.32%
4-Year Graduation, Regents Diploma	0.35	0.48	0	1	68.24%
4-Year Graduation, Advanced Regents Diploma	0.15	0.36	0	1	67.49%
N = 10,000,453					

I do not, however, use all teachers for these empirical exercises. In violation of the homoskedasticity assumption, teachers who appear in the data in more years are lower-variance. Figure 1 shows results from partitioning the data by the number of years each teacher appears in the data, plotted in solid lines in the first panel, and on the whole data, plotted in solid lines in the second panel. Although modified-KS and the likelihood-based estimators give very similar answers when the data is partitioned, they do not give similar answers on the whole data, because they differ in how much more weight they give to teachers who appear in the data more frequently. Modified-KS gives lower estimates because it gives more weight to teachers who appear often in the data.

In order to ensure that the estimates are comparable, all further analysis proceeds on

Figure 1: $\widehat{\text{Var}}(\mu_j)$*: Estimates from partitioning the data by number of years taught, in solid lines, and from whole data, in dotted horizontal lines.*

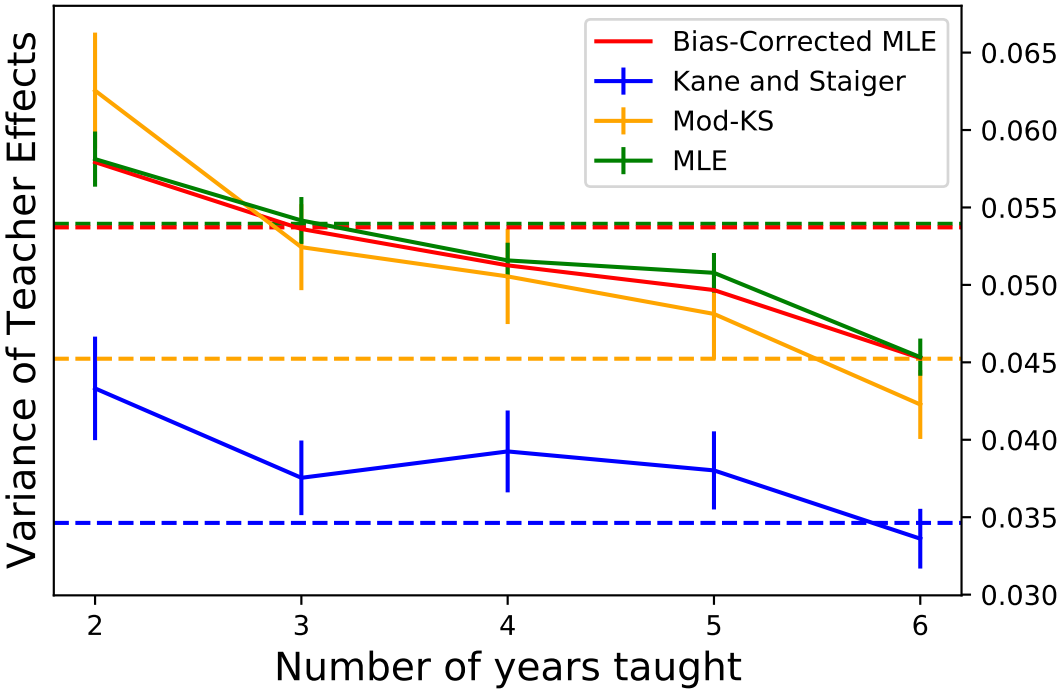

Table 4: *Summary statistics, New York City public schools: students of math teachers who appear in the data for three years.*

	Mean	St. Dev	Min	Max	Missing
Grade	5.69	1.41	4	8	0%
Year	2008.87	2.00	2006	2013	0%
Disabled	0.20	0.40	0	1	0%
Female	0.49	0.50	0	1	0%
English Language Learner	0.13	0.34	0	1	0%
Free Lunch	0.84	0.37	0	1	0%
Days absent	11.96	12.28	0	178	0%
Days present	169.49	13.20	2	186	0%
Days Absent Lag (Z-Score)	0.01	0.93	-12.70	1.08	2.79%
Math Score (Z-Score)	-0.04	0.99	-6.35	3.89	0%
Math score lag (Z-Score)	-0.05	1.00	-10.00	3.96	5.2%
ELA Score (Z-Score)	-0.05	1.00	-10.48	7.76	0%
ELA Score Lag (Z-Score)	-0.05	0.99	-11.10	6.96	8.02%
4-Year Graduation	0.66	0.48	0	1	66.06%
4-Year Grad, Reg	0.44	0.50	0	1	66.06%
4-Year Grad, Adv Reg	0.17	0.38	0	1	66.05%
N = 225,320					

Table 5: *Estimates of $\widehat{\text{Var}}(\mu_j)$ for the effects of math teachers on math scores, for teachers who can be linked to students.*

	$\hat{\sigma}^2_\mu$	$\widehat{\text{Var}}(\bar{\hat{x}}_j^T\lambda)$	$\widehat{\text{Var}}(\mu_j)$	CI	CI (Robust)
Kane and Staiger			3.46%	[3.46%, 3.47%]	
Mod-KS			4.52%	[4.52%, 4.53%]	
MLE	3.15%	2.24%	5.39%	[5.39%, 5.40%]	[5.31%, 5.47%]
Bias-Corrected MLE	3.15%	2.22%	5.37%		
Bias-Corrected MLE (Robust)	3.15%	0.00%	3.15%		

*Notes.*The BC-MLE estimate shows estimates from bias-corrected maximum likelihood using the inverse Fisher information matrix to estimate the variance of $\hat{\lambda}$, which is used to estimate a bias correction; the "sandwich-based" version uses a robust estimator of the variance of $\hat{\lambda}$ to create the bias correction.

Table 6: *Estimates of $\widehat{\text{Var}}(\mu_j)$ for the effects of math teachers on math scores, for teachers who appear in the data in exactly three years and can be linked to students.*

	$\hat{\sigma}^2_\mu$	$\widehat{\text{Var}}(\bar{\bar{x}}_j^T\lambda)$	$\widehat{\text{Var}}(\mu_j)$	CI	CI (Robust)
Kane and Staiger			3.75%	[3.75%, 3.76%]	
Mod-KS			5.24%	[5.24%, 5.25%]	
MLE	3.16%	2.25%	5.42%	[5.41%, 5.42%]	[5.34%, 5.50%]
Bias-Corrected MLE	3.16%	2.20%	5.36%		
Bias-Corrected MLE (Robust)	3.16%	0.02%	3.19%		

*Notes.*The BC-MLE estimate shows estimates from bias-corrected maximum likelihood using the inverse Fisher information matrix to estimate the variance of $\hat{\lambda}$, which is used to estimate a bias correction; the "sandwich-based" version uses a robust estimator of the variance of $\hat{\lambda}$ to create the bias correction.

a sample of teachers who appear in the data in three separate years. This leaves 2,922 teachers out of an initial sample of 25,508. Table 4 gives summary statistics for this reduced dataset. Table 6 shows estimates of $\widehat{\text{Var}}(\mu_j)$ for each estimator for the effects of math teachers on math scores, among teachers who appear in the data in three years. The maximum likelihood estimator has confidence intervals corresponding to both non-robust (inverse Fisher information) and robust to misspecification (sandwich) estimates. Because the bias-corrected maximum likelihood estimator depends on an estimate of the asymptotic variance of $\hat{\lambda}$, there are two possible bias corrections available. The Modified-Kane and Staiger, MLE, and bias-corrected MLE estimators give very similar answers after restricting to teachers who teach in only three years, all between 5.24% and 5.42%. However, the Kane and Staiger estimator gives a much lower answer, 3.75%. This is not surprising, since the Kane and Staiger estimator is not consistent under the baseline model. More surprisingly, the bias-corrected likelihood estimate that relies on a robust estimate of the variance of $\hat{\lambda}$ gives a much smaller answer, because it estimates a much larger variance of $\hat{\lambda}$. Closed-form formulas for confidence intervals are not available for the bias-corrected estimates, since the bias correction depends on the asymptotic variance of $\hat{\lambda}$.

4.2 Subsampling Experiments

To assess the validity of confidence intervals based on asymptotic approximation, I follow Buchinsky (1995) in treating estimates $\widehat{\text{Var}}(\mu_j)$ from the whole sample as the truth, drawing small subsamples b of the data, constructing a nominally 95% confidence interval $\text{CI}_{(b)}$ for $\widehat{\text{Var}}(\mu_j)_{(b)}$ based on the asymptotic distribution of the estimator, and checking how often $\widehat{\text{Var}}(\mu_j)$ lies in the confidence interval. For example, the empirical coverage of estimates from the Kane and Staiger estimator is

$$\widehat{\text{coverage}}_{\text{KS}} = \frac{1}{N_{\text{draws}}} \sum_{b=1}^{N_{\text{draws}}} \mathbb{1}\left(\widehat{\text{Var}}(\mu_j)^{\text{KS}} \in \text{CI}^{\text{KS}}_{(b)}\right),$$

with empirical coverage probabilities constructed similarly for the other estimators: Modified-KS, and MLE with both robust and non-robust confidence intervals.

	Number of teachers	
	200	500
KS	92.0%	96.2%
Mod-KS	88.6%	94.7%
MLE	58.5%	71.2%
MLE (Robust)	100.0%	100.0%

Table 7: *Empirical coverage probabilities of confidence intervals based on asymptotic approximations over 1000 draws.*

Table 7 gives estimated empirical coverage probabilities based on 1000 subsamples of 200 teachers and 1000 subsamples of 500 teachers. Both the KS and modified-KS estimators have confidence intervals that are anti-conservative with 200 teachers but approximately correct with 500 teachers. The quasi-ML estimator using non-robust standard errors produces confidence intervals with very poor coverage. It may struggle because it estimates more parameters than the moment-matching estimators. The robust standard errors, on the other hand, are too large.

In addition to evaluating confidence intervals, comparing the distribution of subsampled estimates $\widehat{\text{Var}}(\mu_j)_{(b)}$ to the "true" answer $\widehat{\text{Var}}(\mu_j)$ allows for empirical estimates of the bias and variance of each estimator, as well as helping visualize the distribution. Figure 2 histograms draws of $\widehat{\text{Var}}(\mu_j)_{(b)}$ from each estimator.

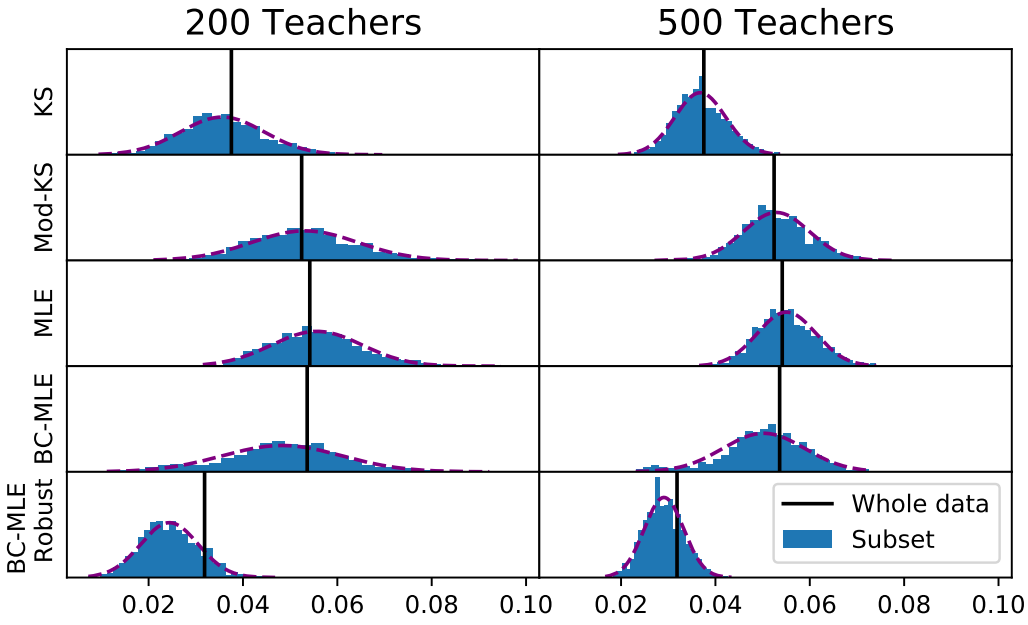

Figure 2: *Histograms of estimates of* $\widehat{\text{Var}}(\mu_j)$ *from subsamples.*

When discussing how estimates from small subsamples relate to the "true" answer from

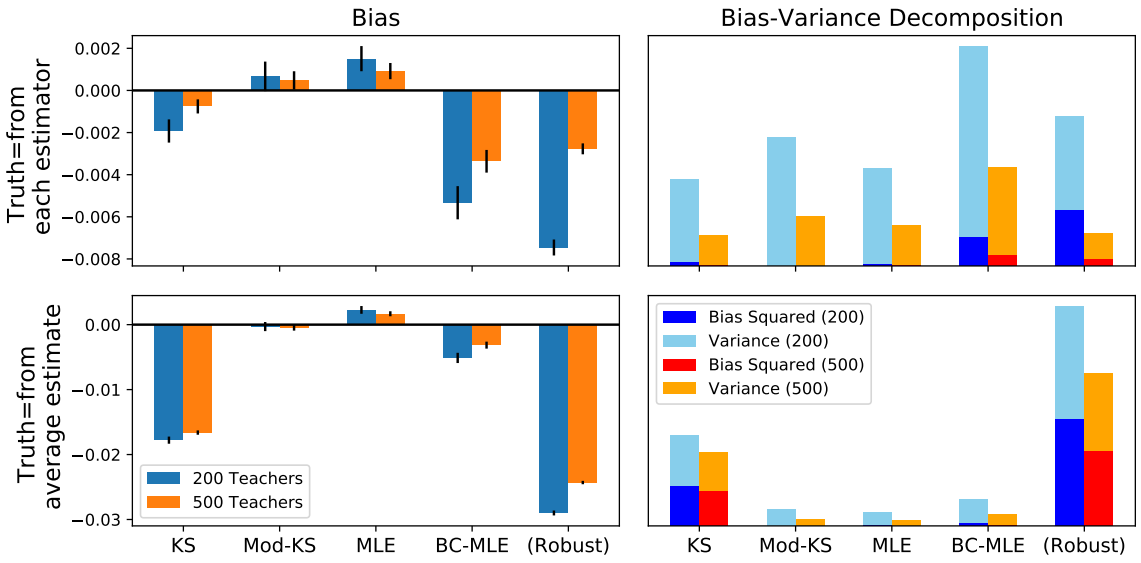

Figure 3: *Bias and variance of each estimator based on subsampling experiments.*

Notes. At left, bias for each estimator, taking estimate from the whole sample as the truth; at right, decomposition of mean squared error into bias squared and variance. At top, defining the "truth" for each estimator as that estimator's estimate from the full sample of all teachers who appear in three years. At bottom, defining the truth as the average estimate from Mod-KS, MLE, and bias-corrected MLE.

the full data, there are two ways of defining the "true" answer: Either the truth is the corresponding full-sample estimate for each estimator, or it is between the "reasonable" estimates of 5.24% to 5.42% from the three consistent estimators that agree with each other. I consider both cases. The top left panel of of Figure 3 shows the "bias" of subsample estimates of $\widehat{\text{Var}}(\mu_j)$, where bias is the mean subsample estimate minus the estimate from the whole sample for that estimator. Even though the Kane and Staiger estimator gives the lowest estimates, it has a small upward "bias" here because it is, on average, close to its value on the whole sample. But the bottom left panel takes the true value to be the mean of the estimates from Modified-KS, MLE, and Bias-Corrected MLE, in which case the Kane and Staiger estimator does much worse. The right panels repeat this exercise with bias squared and variance, which add up to each estimator's mean squared error. It is apparent that the modified-KS is more variable than the Kane and Staiger estimator, but Kane and Staiger's bias makes it unappealing. The likelihood-based estimators perform better in terms of both bias and variance.

5 Conclusion

Each of the estimators presented has pros and cons for estimating the distribution of value-added. 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.*, 2013b). 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 quasi-likelihood is less computationally efficient but appears in subsampling experiments to be more statistically efficient. It is biased upwards, but the bias appears to be quantitatively small in realistic scenarios. A bias correction is available but is not recommended unless an underestimate is strongly preferred to an overestimate, as it overshoots and increases variance with the large number of covariates found in a realistic education example. 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). 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..

When the model is correctly specified, both modified-KS and quasi-ML are appealing choices. However, two complications to the model may influence the choice of estimator: sorting on variables measured with error, and covariates that do not vary within teacher.

When teachers are sorted to students on variables measured with error, modified-KS estimates is more affected by attenuation bias than Kane and Staiger estimates. If the researcher has reason to believe the more stringent identification criteria of Kane and Staiger – $\boldsymbol{\lambda} = 0$ – then the Kane and Staiger estimator will perform better than the modified-KS estimator.

When there are covariates that do not vary within teacher, such as gender or where the teacher went to college, $\widehat{\text{Var}}(\mu_j)$ is not point identified without further assumptions. Although both the modified-KS and quasi-ML estimators give similar estimates when all covariates vary within teacher, quasi-ML is better able to explore the implications of varying the sorting parameter $\boldsymbol{\alpha}$ because it fully models the sorting process. Quasi-ML can generate an estimate of $\widehat{\text{Var}}(\mu_j)$ for $\boldsymbol{\alpha} = (0, \ldots, 0)^T$, $\boldsymbol{\alpha} = (1, \ldots, 1)^T$, and the value of $\boldsymbol{\alpha}$ that minimizes $\widehat{\text{Var}}(\mu_j)$, while mod-KS can only give estimates for $\boldsymbol{\alpha} = (0, \ldots, 0)$ or the value that minimizes $\widehat{\text{Var}}(\mu_j)$. Furthermore, in asymptopia quasi-ML gives a higher lower bound than mod-KS.

References

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

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*.

Buchinsky, M. (1995). Estimating the asymptotic covariance matrix for quantile regression models: A Monte Carlo Study. *Journal of Econometrics*, **68**, 303–338.

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. (2010). 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.

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. 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*.

— and Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Tech. rep., National Bureau of Economic Research.

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 3.1, 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 the parameters $\eta = \left(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, \boldsymbol{\beta}, \boldsymbol{\lambda}\right)$, with true value $\eta^* = \left(\sigma_\mu^{2*}, \sigma_\theta^{2*}, \sigma_\varepsilon^{2*}, \boldsymbol{\beta}^*, \boldsymbol{\lambda}^*\right)$. Consider the normal model

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

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

Lemma A.1.

$$\eta^* = \eta_F \equiv \arg\max_\eta \mathrm{E}_\mathcal{D} \log f\left(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \eta\right) \tag{12}$$

Proof.

$$\begin{aligned}\mathrm{E}_\mathcal{D} \log f(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \eta) = &-\frac{1}{2}\log \det\Sigma(\eta, s_j) \\ &-\frac{1}{2}\mathrm{E}_\mathcal{D}\left[\left(\boldsymbol{y}_j - \boldsymbol{x}_j\boldsymbol{\beta} - \bar{\bar{x}}_j^T\boldsymbol{\lambda}\right)^T \Sigma(\eta, s_j)^{-1}\left(\boldsymbol{y}_j - \boldsymbol{x}_j\boldsymbol{\beta} - \bar{\bar{x}}_j^T\boldsymbol{\lambda}\right) | \boldsymbol{x}_j, \bar{\bar{x}}_j, s_j\right]\end{aligned} \tag{13}$$

Since $\boldsymbol{\beta}$ corresponds to an unrestricted linear predictor, the values of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ that maximize Equation 13 do not depend on Σ, so

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

After plugging in $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$, we can rewrite Equation 13 as

$$
\begin{aligned}
&\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{\beta}^*, \boldsymbol{\lambda}^*) \\
&= -\frac{1}{2} \log \det \Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j) \\
&\quad - \frac{1}{2} \mathrm{E}_{\mathcal{D}} \left[\left(\boldsymbol{y}_j - \boldsymbol{x}_j \boldsymbol{\beta}^* - \bar{\bar{x}}_j^T \boldsymbol{\lambda}^* \right)^T \Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \left(\boldsymbol{y}_j - \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(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j) \\
&\quad - \frac{1}{2} \mathrm{E}_{\mathcal{D}} \left[\mathrm{trace} \left(\boldsymbol{y}_j - \boldsymbol{x}_j \boldsymbol{\beta} - \bar{\bar{x}}_j^T \boldsymbol{\lambda} \right)^T \Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \left(\boldsymbol{y}_j - \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(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j) \\
&\quad - \frac{1}{2} \mathrm{trace} \left(\Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \mathrm{E}_{\mathcal{D}} \left[\left(\boldsymbol{y}_j - \boldsymbol{x}_j \boldsymbol{\beta}^* - \bar{\bar{x}}_j^T \boldsymbol{\lambda}^* \right)^T \left(\boldsymbol{y}_j - \boldsymbol{x}_j \boldsymbol{\beta}^* - \bar{\bar{x}}_j^T \boldsymbol{\lambda}^* \right) \right] \right) \\
&\quad = -\frac{1}{2} \log \det \Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j) - \frac{1}{2} \mathrm{trace} \left(\Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \Sigma(\eta^*, s_j) \right) \\
&\quad \equiv -\frac{1}{2} \log \det \Sigma(\eta, s_j) - \frac{1}{2} \mathrm{trace} \left(\Sigma(\eta, s_j)^{-1} \Sigma(\eta^*, s_j) \right)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\underset{\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2}{\arg\max} \mathrm{E}_{\mathcal{D}} \log f(\boldsymbol{y}_j | \boldsymbol{x}_j, \bar{\bar{x}}_j; \sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, \boldsymbol{\beta}, \boldsymbol{\lambda}) \\
&\quad = \underset{\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2}{\arg\max} - \log \det \Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j) - \mathrm{trace} \left(\Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \Sigma(\eta^*, s_j \right) \\
&\quad = \underset{\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2}{\arg\max} \log \det \left(\Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \Sigma(\eta^*, s_j) \right) - \mathrm{trace} \left(\Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \Sigma(\eta^*, s_j \right)
\end{aligned}
$$

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

$$
\begin{aligned}
&\log \det \left(\Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \Sigma(\eta^*, s_j) \right) - \mathrm{trace} \left(\Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \Sigma(\eta^*, s_j \right) \\
&\quad = \log \det \left(\Sigma(\eta^*, s_j)^{1/2} \Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \Sigma(\eta^*, s_j)^{1/2} \right) \qquad (14) \\
&\qquad - \mathrm{trace} \left(\Sigma(\eta^*, s_j)^{1/2} \Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j)^{-1} \Sigma(\eta^*, s_j)^{1/2} \right) \\
&\quad = \log \Pi_i e_i - \sum_i e_i \\
&\quad = \sum_i \left(\log(e_i) - e_i \right) \qquad (15)
\end{aligned}
$$

Equation 14 is maximized when all $e_i = 1$, which occurs when $\Sigma(\sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, s_j) = \Sigma(\eta^*, s_j)$. As long as the teacher teaches multiple classes and at least one class has multiple students, the only value that solves this equation is $\eta_F = \eta^*$.

□

B Closed-Form Likelihood and Intuitive Parameter Estimates

This section derives a closed-form solution for the likelihood. Subsection B.1 translates Equation 12, 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:

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

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 $\boldsymbol{y}_j$:

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

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 $\mu_j + \Theta_j + \varepsilon_j$ 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:

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

B.2 Solving the Integrals

This section derives a closed-form solution for the likelihood using Equation 20 as a starting point. The quantities that fall out of these equations are generally means or deviations from means, using precision weights.

Define classroom-level means and within-classroom demeaned values for y, and analogues for x:

$$\bar{y}_c \equiv \frac{1}{|I(c)|}\sum_{i\in I(c)} y_i \tag{21}$$

$$\tilde{y}_i \equiv y_i - \bar{y}_{c(i)} \tag{22}$$

$$\tag{23}$$

The precision of the mean classroom error is

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

Define precision-weighted teacher-level means and within-teacher demeaned values for y, and analogues for x:

$$\begin{aligned}\bar{\bar{y}}_j &\equiv \frac{\sum_{c:j(c)=j} h_c \bar{y}_c}{\sum_{c:j(c)=j} h_c} \\ \tilde{\bar{y}}_c &\equiv \bar{y}_c - \bar{\bar{y}}_{j(c)}\end{aligned} \tag{25}$$

Note that, where ϕ is the multivariate normal probability density function, and n_c is the number of students in classroom c,

$$\Pi_{i \in I(c)} \phi\left(y_i; \sigma_\varepsilon^2\right) \propto \sigma_\varepsilon^{2-n_c} \phi\left(\sqrt{\sum_i \tilde{\bar{y}}_i^2}, \sigma^2\right) \phi\left(\bar{y}_c, \sigma_\varepsilon^2 / n_c\right). \tag{26}$$

Equation 26 implies

$$\Pi_{i \in I(c)} \phi(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - \mu - \theta; \sigma_\varepsilon) \propto \sigma_\varepsilon^{2-n_c} \phi\left(\sqrt{\sum_{i \in I(c)} \left(\tilde{\bar{y}}_i - \tilde{\bar{\boldsymbol{x}}}_i^T \boldsymbol{\beta}\right)^2}; \sigma_\varepsilon^2\right) \phi\left(\bar{y}_c - \bar{x}_c^T \boldsymbol{\beta} - \mu - \theta; \sigma_\varepsilon^2 / n_c\right),$$

Also note that the product of the densities of two normal distributions, integrated over a translation of their means, is

$$\int_\mu \phi(\mu - x_1; \sigma_1) \phi(\mu - x_2, \sigma_2) \, \mathrm{d}\mu = \phi\left(x_1 - x_2; \sqrt{\sigma_1^2 + \sigma_2^2}\right). \tag{27}$$

Applying Equation 27,

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

Note the product of n normal densities with different means and variances:

$$\begin{aligned}\Pi_c \phi(\mu_c; \sigma_c) &= \sqrt{\frac{1}{\sum_c 1/\sigma_c^2}} \phi\left(\tilde{\mu}, \boldsymbol{I}\sigma^2\right) \phi\left(\bar{\mu}, \frac{1}{\sum_c 1/\sigma_c^2}\right) \\ &\equiv \sqrt{\frac{1}{\sum_c h_c}} \phi\left(\tilde{\mu}, \boldsymbol{I}(1/h)\right) \phi\left(\bar{\mu}, \frac{1}{\sum_c h_c}\right)\end{aligned} \tag{28}$$

Applying Equation 28,

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

Therefore,

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

Applying Equation 27 again,

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

Equation 20 finally reduces to

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

The log-likelihood is

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

The log-likelihood for all teachers is

$$\sum_j \log f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, \bar{\bar{x}}_j, s_j; \sigma_\mu^2, \sigma_\theta^2, \sigma_\varepsilon^2, \boldsymbol{\beta}, \boldsymbol{\lambda}\right)$$

$$= (N_{\text{classes}} - N_{\text{students}}) \log \sigma_\varepsilon + \frac{1}{2}\sum_c \log h_c - \frac{1}{2}\sum_j \log\left(\sum_{c\in C(j)} h_c\right) - \frac{1}{2}\sum_c \left(\tilde{\bar{y}}_c - \tilde{\bar{x}}_c^T\boldsymbol{\beta}\right)^2 h_c$$

$$- \frac{1}{2\sigma_\varepsilon^2}\sum_i \left(\tilde{y}_i - \tilde{x}_i^T\boldsymbol{\beta}\right)^2 - \frac{1}{2}\sum_j \log\left(\sigma_\mu^2 + \frac{1}{\sum_{c\in C(j)} h_c}\right) - \sum_j \frac{1}{2\left(\sigma_\mu^2 + \frac{1}{\sum_{c\in C(j)} h_c}\right)}\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T\left(\boldsymbol{\beta}+\boldsymbol{\lambda}\right)\right)^2.$$

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

$$f\left(\boldsymbol{y}_j|\boldsymbol{x}_j, s_j; \eta\right) = g\left(\eta\right) \int_\mu \phi\left(\mu - \bar{\bar{x}}_j^T\boldsymbol{\lambda}; \sigma_\mu^2\right)\phi\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T\boldsymbol{\beta} - \mu; \frac{1}{\sum_c h_c}\right)\mathrm{d}\mu \tag{29}$$

C Maximum Likelihood Bias Correction

Proof of Equation 6:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\hat{\boldsymbol{\lambda}}\right)\right] - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) &= \mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\hat{\boldsymbol{\lambda}}\right]\right) - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\
&= \mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\right]\hat{\boldsymbol{\lambda}}\right) - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\
&= \mathbb{E}\left[\mathrm{Var}\left(\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\bar{\bar{x}}\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[\bar{\bar{x}}^T\hat{\boldsymbol{\lambda}}|\bar{\bar{x}}\right]\right) - \mathbb{E}\left[\bar{\bar{x}}\right]^T\mathrm{Cov}(\hat{\boldsymbol{\lambda}})\,\mathbb{E}\left[\bar{\bar{x}}\right] - \mathrm{Var}\left(\bar{\bar{x}}^T\boldsymbol{\lambda}\right) \\
&= \mathbb{E}\left[\bar{\bar{x}}^T\,\mathrm{Cov}\left(\hat{\boldsymbol{\lambda}}\right)\bar{\bar{x}}\right] - \mathbb{E}\left[\bar{\bar{x}}\right]^T\mathrm{Cov}(\hat{\boldsymbol{\lambda}})\,\mathbb{E}\left[\bar{\bar{x}}\right] \\
&= \mathbb{E}\left[\left(\bar{\bar{x}} - \mathbb{E}\,\bar{\bar{x}}\right)^T\mathrm{Cov}\left(\hat{\boldsymbol{\lambda}}\right)\left(\bar{\bar{x}} - \mathbb{E}\,\bar{\bar{x}}\right)\right].
\end{aligned} \tag{30}$$

D Optimization,Gradient

This equation can easily be optimized numerically. The software package available at http://www.github.com/esantorella/tva iterates between estimating $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}$, 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 \left(\boldsymbol{\beta} + \boldsymbol{\lambda} \right) \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 \left(\boldsymbol{\beta} + \boldsymbol{\lambda} \right) \right)^2 \right) \\
& - \sum_j \frac{\frac{1}{\sum_{c \in C(j)} h_c}}{\sigma_\mu^2 + \frac{1}{\sum_{c \in C(j)} h_c}} \left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T \left(\boldsymbol{\beta} + \boldsymbol{\lambda} \right) \right) \left(\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T (\beta + \lambda) \right) \sum_{c \in C(j)} h_c^2 - \sum_{c \in C(j)} h_c^2 (\bar{y}_c - \bar{x}_c^T (\beta + \lambda)) \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 \left(\boldsymbol{\beta} + \boldsymbol{\lambda} \right) \right)^2 \right) \\
& - \sum_j \frac{\frac{1}{\sum_{c \in C(j)} h_c}}{\sigma_\mu^2 + \frac{1}{\sum_{c \in C(j)} h_c}} \left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T \left(\boldsymbol{\beta} + \boldsymbol{\lambda} \right) \right) \left(\left(\bar{\bar{y}}_j - \bar{\bar{x}}_j^T (\beta + \lambda) \right) \sum_{c \in C(j)} h_c^2 / n_c - \sum_{c \in C(j)} h_c^2 / n_c (\bar{y}_c - \bar{x}_c^T (\beta + \lambda)) \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 \left(\boldsymbol{\beta} + \boldsymbol{\lambda} \right) \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 \left(\boldsymbol{\beta} + \boldsymbol{\lambda} \right) \right) \bar{\bar{x}}_j
\end{aligned}
$$

E Bounding the Asymptotic Bias in the Kane and Staiger Procedure

We know that

$$\begin{aligned}
\text{Bias}(\widehat{\text{Var}}(\mu)_j) =& -2\frac{2}{\sum_j |C(j)|\,|C(j)-1|}\,\mathbb{E}\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T \sum_{c,c'\in C(j)} \bar{\boldsymbol{x}}_c \mu_j\right] \\
&+ \frac{2}{\sum_j |C(j)|\,|C(j)-1|}\,\mathbb{E}\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T \left(\sum_{c,c'\in C(j)} \bar{x}_c \bar{x}_{c'}^T\right)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right] \\
=& -2\left(\sum_i \mu_{j(i)} x_i^T\right)\left(\sum_i x_i x_i^T\right)^{-1}\left(\frac{2}{\sum_j |C(j)|\,|C(j)-1|}\sum_{c,c'\in C(j)} \bar{\boldsymbol{x}}_c \mu_j\right) \\
&+ \left(\sum_i \mu_{j(i)} x_i^T\right)\left(\sum_i x_i x_i^T\right)^{-1}\left(\frac{2}{\sum_j |C(j)|\,|C(j)-1|}\sum_{c,c'\in C(j)} \bar{x}_c \bar{x}_{c'}^T\right)\left(\sum_i x_i x_i^T\right)^{-1}\left(\sum_i \mu_{j(i)} x_i\right) \\
=& \,\text{Var}(\mu_j) - 2\left(\sum_i \mu_{j(i)} x_i^T\right)\left(\sum_i x_i x_i^T\right)^{-1}\left(\frac{2}{\sum_j |C(j)|\,|C(j)-1|}\sum_{c,c'\in C(j)} \bar{\boldsymbol{x}}_c \mu_j\right) \\
&+ \left(\sum_i \mu_{j(i)} x_i^T\right)\left(\sum_i x_i x_i^T\right)^{-1}\left(\frac{2}{\sum_j |C(j)|\,|C(j)-1|}\sum_{c,c'\in C(j)} \bar{x}_c \bar{x}_{c'}^T\right)\left(\sum_i x_i x_i^T\right)^{-1}\left(\sum_i \mu_{j(i)} x_i\right)
\end{aligned}$$

In the special case where each teacher teaches in the same number of classrooms and each classroom has the same number of students, this simplifies to

$$\text{Bias}(\widehat{\text{Var}}(\mu_j)) = -2\left(\frac{\sum_j \mu_j \bar{\bar{x}}_j^T}{N_{\text{teachers}}}\right)\left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}\right)^{-1}\left(\frac{\sum_j \mu_j \bar{\bar{x}}_j}{N_{\text{teachers}}}\right) \tag{31}$$

$$+\left(\frac{\sum_j \mu_j \bar{\bar{x}}_j^T}{N_{\text{teachers}}}\right)\left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}\right)^{-1}\left(\frac{\sum_{c,c'\in C(j)} \bar{x}_c \bar{x}_{c'}^T}{N_{\text{classroompairs}}}\right)\left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}\right)^{-1}\left(\frac{\sum_j \mu_j \bar{\bar{x}}_j}{N_{\text{teachers}}}\right) \tag{32}$$

We want to show that

$$-b^T\left(\frac{1}{N_{\text{students}}}\sum_i x_i x_i^T\right) b \le b\frac{1}{N_{\text{class pairs}}}\sum_j \sum_{c,c'\in C(j)} \bar{x}_c \bar{x}_{c'}^T \le b^T\left(\frac{1}{N_{\text{students}}}\sum_i x_i x_i^T\right) b \quad \forall b \in \mathcal{R}^k. \tag{33}$$

Note that

$$\begin{aligned}
\frac{1}{N_{\text{students}}}\sum_i x_i x_i^T &= \frac{1}{N_{\text{students}}}\sum_i \left(\bar{\tilde{x}}_{j(i)} + \tilde{\tilde{x}}_{c(i)} + \tilde{x}_i\right)\left(\bar{\tilde{x}}_{j(i)} + \tilde{\tilde{x}}_{c(i)} + \tilde{x}_i\right)^T \\
&= \frac{1}{N_{\text{teachers}}}\sum_j \bar{\tilde{x}}_j \bar{\tilde{x}}_j^T + \frac{1}{N_{\text{classrooms}}}\sum_c \tilde{\tilde{x}}_c \tilde{\tilde{x}}_c^T + \frac{1}{N_{\text{students}}}\sum_i \tilde{x}_i \tilde{x}_i \\
\frac{1}{N_{\text{class pairs}}}\sum_j \sum_{c,c' \in C(j)} \bar{x}_c \bar{x}_{c'}^T &= \frac{1}{N_{\text{class pairs}}}\sum_j \sum_{c,c' \in C(j)} \left(\bar{\tilde{x}}_j + \tilde{\tilde{x}}_c\right)\left(\bar{\tilde{x}}_j + \tilde{\tilde{x}}_{c'}\right) \\
&= \frac{1}{N_{\text{teachers}}}\bar{\tilde{x}}_j \bar{\tilde{x}}_j^T + \frac{1}{N_{\text{teachers}}}\frac{1}{|C(j)|(|C(j)|-1)}\sum_j \sum_{c,c' \in C(j)} \tilde{\tilde{x}}_c \tilde{\tilde{x}}_{c'}^T \\
&= \frac{1}{N_{\text{teachers}}}\bar{\tilde{x}}_j \bar{\tilde{x}}_j^T - \frac{1}{N_{\text{classrooms}}}\frac{1}{|C(j)|-1}\sum_c \tilde{\tilde{x}}_c \tilde{\tilde{x}}_c^T
\end{aligned}$$

$$\begin{aligned}
\frac{1}{N_{\text{students}}}\sum_i x_i x_i^T - \frac{1}{N_{\text{class pairs}}}\sum_j \sum_{c,c' \in C(j)} \bar{x}_c \bar{x}_{c'}^T &= \frac{1}{N_{\text{classrooms}}}\sum_c \tilde{\tilde{x}}_c \tilde{\tilde{x}}_c^T \left(1 + \frac{1}{|C(j)|-1}\right) + \frac{1}{N_{\text{students}}}\sum_i \tilde{x}_i \tilde{x}_i^T \\
\frac{1}{N_{\text{students}}}\sum_i x_i x_i^T + \frac{1}{N_{\text{class pairs}}}\sum_j \sum_{c,c' \in C(j)} \bar{x}_c \bar{x}_{c'}^T &= \frac{1}{N_{\text{classrooms}}}\sum_c \tilde{\tilde{x}}_c \tilde{\tilde{x}}_c^T \frac{|C(j)|}{|C(j)|-1} + \frac{1}{N_{\text{students}}}\sum_i \tilde{x}_i \tilde{x}_i^T
\end{aligned} \tag{34}$$

Both the right hand side of both equations in Equation 34 are sums of positive definite matrices, proving Equation 33. Therefore, we know that

$$\begin{aligned}
&\left|\left(\frac{\sum_j \mu_j \bar{\tilde{x}}_j^T}{N_{\text{teachers}}}\right)\left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}\right)^{-1}\left(\frac{\sum_{c,c' \in C(j)} \bar{x}_c \bar{x}_{c'}^T}{N_{\text{classroom pairs}}}\right)\left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}\right)^{-1}\left(\frac{\sum_j \mu_j \bar{\tilde{x}}_j}{N_{\text{teachers}}}\right)\right| \\
&\qquad \leq \left|\left(\frac{\sum_j \mu_j \bar{\tilde{x}}_j^T}{N_{\text{teachers}}}\right)\left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}\right)^{-1}\left(\frac{\sum_j \mu_j \bar{\tilde{x}}_j}{N_{\text{teachers}}}\right)\right|
\end{aligned}$$

Plugging this into Equation 31, we find that

$$\begin{aligned}
\text{Bias}(\widehat{\text{Var}}(\mu_j)) &\leq -\left(\frac{\sum_j \mu_j \bar{\tilde{x}}_j^T}{N_{\text{teachers}}}\right)\left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}\right)^{-1}\left(\frac{\sum_j \mu_j \bar{\tilde{x}}_j}{N_{\text{teachers}}}\right) \\
\text{Bias}(\widehat{\text{Var}}(\mu_j)) &\geq -3\left(\frac{\sum_j \mu_j \bar{\tilde{x}}_j^T}{N_{\text{teachers}}}\right)\left(\frac{\sum_i x_i x_i^T}{N_{\text{students}}}\right)^{-1}\left(\frac{\sum_j \mu_j \bar{\tilde{x}}_j}{N_{\text{teachers}}}\right).
\end{aligned}$$