Adding Value to Value-Added: Theory and Applications to Teachers and Bureaucrats

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:40049975

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Adding Value to Value-Added: Theory and Applications to Teachers and Bureaucrats

A dissertation presented
by
Elizabeth Santorella
to
The Department of Economics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Economics

Harvard University
Cambridge, Massachusetts
January 2018
Abstract

Value-added estimators are extensively used to study teachers and other groups. In addition to estimating an individual’s causal impact on outcomes, value-added models describe the dispersion of individual effects: For example, variation in teacher quality contributes about 2% of the variance in their students’ same-year test scores. In this dissertation, I explore the use of value-added estimators in different contexts and discuss the statistical properties of these estimators.

In the first chapter, I model teacher value-added as a vector encompassing teacher effects on many outcomes. The covariance matrix of teacher value-added reveals that teachers have large effects on high school graduation and future test scores, but that immediately-observable teacher effects are poor proxies for future effects. Teacher effects on attendance are about as persistent and as predictive of future achievement as teacher effects on test scores. In the second chapter, I discuss statistical issues that arise in finite samples. In the third chapter, joint with Jonas Hjort and Gautam Rao, I take an in-depth look at one context in which relatively small data induces finite-sample biases: the effects of bureaucrats on local economic outcomes in India. Although point estimates suggest that bureaucrat quality is an important determinant of variation in night light intensity and project completion, randomization inference shows that point estimates are biased upwards and insignificant.
# Contents

Abstract ................................................................. iii  
Acknowledgments ......................................................... x  

## Introduction 1

### 1 Multi-Dimensional Teacher Effects 4

1.1 Literature .......................................................... 10  
1.2 Model ............................................................... 11  
1.3 Data, Setting, and Descriptive Statistics 12

1.3.1 Descriptive Statistics ............................................. 14  
1.4 Estimation ............................................................ 24

1.4.1 Inference .......................................................... 24  
1.4.2 Identification ..................................................... 25  
1.5 Results ............................................................... 26

1.5.1 Are current teacher effects a good proxy for future teacher effects? 33  
1.5.2 Bivariate Correlations and Factor Analysis .................. 34  
1.6 Robustness Checks .................................................. 38  
1.7 Conclusion ............................................................ 42  

## 2 Measuring the Distribution of (Teacher) Value-Added 45

2.1 Toy Model, Motivation ............................................... 48  
2.2 Literature ............................................................ 48  
2.3 Estimators ............................................................. 52

2.3.1 Maximum (Quasi-)Likelihood .................................... 54  
2.3.2 Empirical Bayes estimator from Kane and Staiger (2008) ........ 57  
2.3.3 Modified version of above estimator: Modified-KS ............ 62  
2.3.4 Covariates without within-teacher variation .................... 63  
2.3.5 Errors in variables ............................................... 68  
2.4 Monte Carlo Experiments ............................................ 70

2.4.1 Data and real results ............................................. 70  
2.4.2 Subsampling Experiments ........................................ 74
List of Tables

1.1  Student summary statistics, for whole sample. A grade level of -1 is pre-
    kindergarten. ............................................................. 14
1.2  Student summary statistics, for students who can be matched to teachers.  15
1.3  Predictors of English Language Arts scores. .................................. 18
1.4  Predictors of math scores. ..................................................... 19
1.5  Predictors of attendance. ..................................................... 20
1.6  Predictors of graduation. ..................................................... 21
1.7  Magnitude of math teacher effects on test scores and attendance. .......... 26
1.8  Magnitude of math teacher effects on high school graduation. .............. 27
1.9  Fade out of math teacher effects. .......................................... 30
1.10 Goodness of proxy. .......................................................... 34
1.11 Correlations between teachers’ value-added on different outcomes. ....... 35
1.12 Standard deviation of math teacher effects on ELA scores: Robustness to
    choice of controls and estimator. ....................................... 41
1.13 Standard deviation of math teacher effects on math scores: Robustness to
    choice of controls and estimator. ....................................... 41
1.14 Standard deviation of math teacher effects on attendance: Robustness to
    choice of controls and estimator. ....................................... 41
1.15 Standard deviation of math teacher effects on graduation: Robustness to
    choice of controls and estimator. ....................................... 42
2.1  Estimates of the variance of teacher effects, \( \text{Var}(\mu_j) \), 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. ............... 50
2.2  Comparison of estimators. Asymptotics are as the number of teachers ap-
   proaches infinity. .......................................................... 52
2.3  Summary statistics, New York City public schools. ............................. 70
2.4  Summary statistics, New York City public schools: students of math teachers
    who appear in the data for three years. .................................. 72
2.5 Estimates of $\text{Var}(\mu_j)$ for the effects of math teachers on math scores, for teachers who can be linked to students. 73

2.6 Estimates of $\text{Var}(\mu_j)$ 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. 73

2.7 Empirical coverage probabilities of confidence intervals based on asymptotic approximations over 1000 draws. 75

3.1 Summary statistics for the whole sample and for the largest connected set where each outcome is not null. 83

3.2 Contribution of District Collectors to variation in night light intensity. 94

3.3 Contribution of District Collectors to variation in project completions. 96

3.4 Contribution of District Collectors to variation in placebo outcome. 96

A.1 Magnitude of ELA teacher effects on test scores and attendance. 105

A.2 Magnitude of ELA teacher effects on graduation. 105

A.3 Fade-out of ELA teacher effects. 107

A.4 Goodness of proxy for ELA teachers. 107

A.5 Standard deviation of ELA teacher effects on ELA scores: Robustness to choice of controls and estimator. 108

A.6 Standard deviation of English teacher effects on math scores: Robustness to choice of controls and estimator. 109

A.7 Standard deviation of English teacher effects on attendance: Robustness to choice of controls and estimator. 110

A.8 Standard deviation of English teacher effects on graduation: Robustness to choice of controls and estimator. 111
# List of Figures

1.1 Average number of days absent by ethnicity. ........................................ 16
1.2 Average number of days absent by ethnicity. ........................................ 17
1.3 Binned scatter plots: Relationship between graduation, attendance, and test scores. ................................................................. 22
1.4 Magnitude of math teacher effects on test scores and attendance. .............. 27
1.5 Fade out of math teacher effects. ............................................................ 29
1.6 Magnitude of math teacher effects by grade. ............................................. 31
1.7 Fade out of math teacher effects, by grade. ............................................. 32
1.8 Eigenvalues of $\Sigma_m$ for math teachers. ................................................. 36
1.9 Factors of $\Sigma_m$ for math teachers. ....................................................... 37
1.10 Robustness to choice of estimator. ......................................................... 39
1.11 Robustness to choice of controls. ........................................................... 40
2.1 $\text{Var}(\mu_j)$: Estimates from partitioning the data by number of years taught, in solid lines, and from whole data, in dotted horizontal lines. ............... 71
2.2 Histograms of estimates of $\text{Var}(\mu_j)$ from subsamples. ...................... 75
2.3 Bias and variance of each estimator based on subsampling experiments. .... 76
3.1 Histogram of the number of project completions per district-month. ........... 84
3.2 Histogram of the number of project completions per posting. ................... 85
3.3 Night light intensity z-scores over time, in the seven biggest states and in all of India. ................................................................. 85
3.4 Distribution of simulation estimates for the contribution of District Collectors to variation in night light intensity. From the second specification and Mod-KS estimator. ................................................................. 95
3.5 Distribution of simulation estimates for the contribution of District Collectors to variation in number of project completions. From the second specification and Modified-KS estimator. ............................................. 97
3.6 Distribution of simulation estimates for the contribution of District Collectors to variation in the placebo outcome. From the second specification and Mod-KS estimator. .................................................... 98
| A.1  | Magnitude of ELA teacher effects. | 104 |
| A.2  | Fade-out of ELA teacher effects.  | 106 |
| A.3  | Magnitude of ELA teacher effects by grade. | 108 |
| A.4  | Fade-out of ELA teacher effects by grade. | 109 |
| A.5  | Eigenvalues of $\Sigma_{\mu}$ for English teachers. | 110 |
| A.6  | Factors of $\Sigma_{\mu}$ for ELA teachers. | 111 |
| A.7  | Robustness of ELA teacher effects to estimator. | 112 |
| A.8  | Robustness of ELA teacher effects to choice of controls. | 113 |
Acknowledgments

I am extremely grateful for comments and suggestions from my advisers, Gautam Rao, Gary Chamberlain, and Elie Tamer, to seminar participants in the Harvard University Econometrics and Public Economics lunches, and to many others, including Raj Chetty, Christopher Jencks, Max Kasy, Larry Katz, and Nathan Hendren.

I am indebted to the New York City Department of Education for assistance in obtaining and using the data used in the first chapter.

Many of the computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.
Introduction

Value-added estimators have been extensively used to study teachers and other groups. These estimators describe how dispersed individuals are in their effects on an outcome: for example, variation in teacher quality contributes to about 2% of the variance in student test scores, so a teacher one standard deviation above average improves her students’ test scores by $\sqrt{0.02} = 0.14$ standard deviations. Value-added modeling has been used widely to study the distribution of teacher effects on test scores and is used by many school districts to rank and evaluate teachers. In this dissertation, I explore the use of value-added estimators in different contexts and discuss the statistical properties of these estimators.

I think of a value-added model as one with the following properties: Each observation $i$ corresponds to some entity $j(i)$, where each $j$ corresponds to many $i$. For example, $j(i)$ may be the teacher of student $i$, or $j(i)$ may be the doctor who treats patient $i$. Each teacher $j$ has an average causal effect or “value-added” $\mu_j$. The $\mu_j$ are drawn identically and independently from the same distribution, $\mu_j \sim F$. Although the individual effects $\mu_j$ are often of interest, I focus on estimating the distribution $F$. The $\mu_j$ may be correlated with other factors that influence the outcome, making high-dimensional covariates a common complication.

In the first chapter, “Multi-Dimensional Teacher Effects,” I use data from New York City public schools to estimate the covariance structure of teacher effects on several outcomes: present and future test scores, present and future attendance, and high school graduation. The covariance matrix of teacher effects reveals the magnitude of teacher effects on each outcome and the relationship between teacher effects on different outcomes while sidestepping the need to estimate individual teacher effects. This analysis shows that although teachers
have large effects on longer-term outcomes, these effects are not captured well by short-term effects on test scores or on attendance, which fade out quickly.

Teachers have substantial effects on high school graduation, and on test scores and attendance four years in the future. Students of a teacher who is one standard deviation above average at improving graduation rates are 5 percentage points more likely to graduate high school. Although teacher quality is an important determinant of graduation rates and of future test scores and attendance, long-term effects cannot be predicted well by short-term effects. For example, even if teacher effects on contemporaneous outcomes were perfectly measured, they would only explain about 3% of the variance in teacher effects on high school graduation. My results also suggest that teacher effects on attendance could be an important supplement to score-based measures of teacher value-added. Teachers who improve attendance tend to improve high school graduation rates. However, teacher effects on attendance are only weakly correlated with effects on test scores. I present extensive descriptive evidence on the importance of attendance in the New York City schools. I also show that teachers appear to have smaller but more persistent effects on the test scores of older students. Finally, I use factor analysis to break the correlation matrix of teacher effects into three components. The results are robust to choice of value-added estimator and to choice of control variables.

As value-added estimation spreads to fields outside education, where data sets may be small and experimental validation infeasible, estimators that perform well without millions of observations are increasingly needed. In my second chapter, “Measuring the Distribution of (Teacher) Value-Added,” I study the statistical properties of existing value-added methodologies and develop a new likelihood-based framework. I clarify conditions under which existing methods are identified, sign their biases, and derive asymptotic standard errors; and I develop a likelihood-based estimator.

Subsampling-based Monte Carlo experiments confirm the theoretical predictions about estimators’ biases and indicate that the likelihood-based estimator is slightly more efficient than existing methods. I use these subsampling intervals to estimate the coverage of
confidence intervals based on the asymptotic distribution of the estimator. I also discuss the
effect of errors in variables, and the failure of point identification when some covariates do
not vary within teacher. I conclude by providing advice on when to use which estimator.

In the third chapter, I, with Jonas Hjort and Gautam Rao, use several value-added esti-
mators to study a question relevant to political economy: How much agency do individual
bureaucrats have to impact local economic performance? We study high-ranking bureau-
crats in the Indian Administrative Service, India’s national bureaucracy. These bureaucrats,
District Collectors, are quasi-randomly assigned to manage the bureaucracy of an Indian
district and often transfer to different districts in the same state. This setting presents
econometric challenges, since we have relatively few observations and high-dimensional
covariates. By randomly permuting bureaucrat names in a way consistent with the actual
assignment mechanism, we show that value-added estimators have economically and statis-
tically meaningful finite-sample biases in this setting. Point estimates suggest that variance
in District Collector quality accounts for substantial variance in project completion and
night light intensity. However, randomization inference and bootstrap hypothesis tests show
that our estimates are in fact insignificant.

Finally, the appendix includes proofs and supplementary figures.
Chapter 1

Multi-Dimensional Teacher Effects

Teacher quality is an important determinant of academic achievement (Chetty et al. 2014a, Kane and Staiger 2008, Hanushek and Rivkin 2006). Many states require schools to use test score-based value-added measures in teacher evaluation. However, measures of teacher quality based on short-run test scores may be highly incomplete, because they neglect teacher effects on non-test score measures and on long-term outcomes. Education influences students’ “non-cognitive skills” or “character skills”; these skills, important in life and the labor market, are not well-captured by test scores (Carneiro et al. 2007). If teachers influence their students in ways that are not reflected in test scores, score-based value-added measures miss important components of teacher quality. Furthermore, if teacher effects on short-run outcomes are poor proxies for teacher effects on long-run outcomes, value-added measures based on short-term measurements will be highly incomplete. In this paper, I estimate the variance of teacher effects on students’ test scores and attendance, both contemporaneously and up to four years in the future, and on high school graduation. I estimate the covariance structure of teacher effects on various outcomes and, using these covariances, investigate the relationships between teacher effects in different domains; how quickly the effects of teachers who improve a short-term outcome fade out; and how well short-term teacher effects predict long term teacher effects.

Using administrative data from the New York City public elementary and middle schools,
I find that teachers’ causal effects on attendance — their “attendance value-added” — are highly variable; a teacher whose attendance value-added is one standard deviation above average improves her students’ attendance by about 0.07 standard deviations, or 1.5 days. The correlation between a teacher’s effect on test scores and her effect on attendance is only about 0.12 to 0.15, implying that teacher quality metrics that use only test scores miss much of a teacher’s impact on her students. Much like teacher effects on test scores, these effects fade out quickly: students of a teacher who improves test scores or attendance have test scores or attendance four years later that are barely better than would be predicted by demographic characteristics.

However, teacher quality is an important determinant of high school graduation, future test scores, and future attendance. The presence of both substantial long-term effects and of effects that fade out quickly is not a paradox: the teachers who improve test scores in the long term are not particularly likely to be the teachers who improve test scores in the short term. The lack of similarity between teacher effects on short-term outcomes and on long-term outcomes highlights the presence of a multi-tasking problem: if improving easily-observed outcomes like same-year test scores and attendance is a very different from improving more welfare-relevant outcomes like graduation, incentivizing the more easily observed outcomes may have perverse effects. A teacher who is one standard deviation above average at improving her students’ high school graduation rates increases graduation rates by 5 percentage points and graduation with an Advanced Regents designation by 6 percentage points, but only 3% of this variation can be predicted using immediately-available data (same-year test scores and attendance).

Methodologically, I build on the teacher value-added literature by extending conventional methods to treat value-added as a vector rather than a scalar. Teacher value-added methods typically estimate both the variance of teachers’ causal effects on their students’ outcomes and an individual causal effect, or “value-added”, for each teacher; my method sidesteps

---

1By contrast, the correlation between teacher effects on math test scores and on English Language Arts test scores is 0.66. See Table 1.11.
the need to estimate individual value-added scores. My model is very similar to other value-added models in using variance decompositions and “moment-matching” to estimate what portion of variance in outcomes is due to teachers. These models must avoid giving teachers credit for receiving more able students rather than for their causal effects on test scores. They typically achieve this by controlling for a rich set of student covariates, such as demographic factors and previous test scores; thus, these models estimate the value a teacher adds to a student above what that student would achieve with an average teacher. This literature typically finds that teachers vary moderately to largely in their effects on students: teachers account for about 2% of the variance in test scores. In other words, a teacher who is one standard deviation above average in her effectiveness at increasing test scores (“score VA”) increases her students’ test scores by an average of 0.1 standard deviations. I extend this methodology by incorporating a variety of different outcomes, including leads of outcomes, and estimating not only the variance of teacher effects but the covariance of effects on different measures.

Value-added measures based on short-term test scores have become a common component of teacher evaluations, but there are reasons to suspect that these measures are incomplete, motivating a focus on other aspects of teacher effects. Chetty et al. (2014b) shows that students of teachers who improve test scores are more likely to go to college, have higher incomes, and live in better neighborhoods as adults, but these effects appear to be too large to be explicable by the increase in academic achievement implied by test score gains. And studies that do not specifically involve teachers find that quantity and quality of education improve skills that are not captured by test scores, but the mechanisms for this are unclear.

Rewarding teachers based on test scores is unpopular with many parents, partially due to concerns that test scores reflect only part of teachers’ beneficial effects on their students. If so, policymakers face a multitasking problem in the spirit of Holmstrom and

---

269% of New York State parents say that teacher pay “should not be based on how well their students perform on standardized tests,” while 26% say it should be noa (2015).
Milgrom (1991): we want teachers to make their students motivated, persistent, creative, and informed, but we can’t measure those characteristics. Teachers who are incentivized to increase test scores may behave in counterproductive ways. This concern has empirical merit: Teachers or administrators under low to moderate incentives to improve test scores cheat or manipulate scores (Jacob and Levitt (2003), Dee et al. (2016), Loughran and Comiskey (1999)), spend less time on non-tested subjects (Jacob, 2005), spend much more time on test preparation (Klein et al. (2000)), and move students into special education so that those students will not be counted in school progress indicators (Figlio and Getzler (2002), Jacob (2005)).

One area that could be given short shrift by an increasing focus on test scores is non-cognitive, or character, skills. Education is important in transmitting these skills, such as the drive and persistence to attend school and work hard, and while the mechanisms are poorly understood, teachers may be an important factor. In addition to test scores, I include two outcomes that plausibly reflect non-cognitive skills: attendance and high school graduation. Recent papers show that teachers influence such outcomes in the short term. Gershenson (2016) studies third through fifth graders in North Carolina and finds that teachers have “arguably causal, statistically significant effects on student absences that persist over time,” and that “teachers who improve test scores do not necessarily improve student attendance.” Similarly, Jackson (2016) studies ninth graders in North Carolina and finds that teachers have medium-term effects on student absences, suspensions, grades, and on-time grade progression.

My results replicate those of Jackson (2016) for teacher effects on test scores, attendance, and their relationship to graduation, where applicable. Jackson finds that teacher effects on test scores correlate with effects on a “behavioral factor” at 0.16, and I find that effects on test scores correlate with effects on attendance at 0.12 to 0.15. Jackson finds that increasing teacher value-added on test scores by one standard deviation increases high school graduation by 0.13 percentage points, and a one standard deviation increase in the behavioral factor increases graduation by 0.78 percentage points. I find that increasing
value-added on test scores does not have a statistically significant effect on graduation, and increasing value-added on attendance by one standard deviation would increase graduation by 0.72 percentage points. We both find that using attendance and test scores instead of just test scores to predict teacher effects on graduation more than doubles predictive validity.

However, I emphasize that the ability to predict longer-term outcomes like graduation using short-term outcomes is quite low. Even if perfectly measured, teacher effects on attendance and test scores could only explain 3% of the variance in teacher effects on graduation (and 6% of the variation in graduating with an Advanced Regents designation). Raising teachers’ “attendance value-added” by one standard deviation, and leaving the covariance of teacher effects on other outcomes intact, would boost graduation rates by 0.72 percentage points, while raising “graduation value-added” by one standard deviation would increase graduation rates by 5 percentage points. A similar result appears in Chamberlain (2013), who finds that teacher effects on test scores capture less than 20% of teacher effects on college graduation.

Recent policy changes hint at a shift in focus away from test scores and towards quantitative evaluations that incorporate other metrics. The federal Every Student Succeeds Act of 2015 (ESSA) mandates that each state measure school quality based on a metric of its own devising that includes test scores but also includes at least one substantially different outcome. Many states have chosen to measure and reward attendance.

In this paper, I demonstrate that teachers influence medium-term outcomes (high school graduation and test scores four years ahead), that immediate effects on test scores are a poor proxy for medium-term effects, and that teacher effects on attendance are slightly smaller and about as persistent as teacher effects on test scores.

My estimates are only credible insofar as identification restrictions are satisfied. Teachers must be sorted to students only on observables, meaning that conditional on covariates, no teacher should be systematically assigned students who have unobservable characteristics that cause high or low performance. The rich, longitudinal nature of my data makes it possible to control for a variety of student and classroom characteristics and lagged values.
of outcomes, making the sorting on observables requirement plausible. For example, it is possible that high-SES students or students who have been improving relative to their peers are, on average, assigned to better teachers. But since I observe and control for ethnicity, free lunch status, lagged values of test scores and attendance, and class-level means of these variables, this sorting would be predictable from the control variables and would not violate the sorting on observables restriction. I use a pre-trend test as an empirical check of this restriction. Although other authors have found that controlling for lagged scores is sufficient to find unbiased measures of parameter estimates (Chetty et al., 2014a), I measure a small but statistically significant pre-trend in which teacher effects on contemporaneous outcomes “predict” past achievement. Results are robust to controlling for more lags, including higher-order interactions, and using a different estimator.

This paper also provides descriptive evidence on patterns of absenteeism. The patterns documented are consistent with poor and minority students often missing school voluntarily or for reasons other than illness. Students in New York City are absent extremely often, and chronic absenteeism — missing more than 10% of a school year — is common. Students are far more likely to be absent in later grades, and there are large ethnic gaps in school attendance.

This paper proceeds as follows. In Section 1.1 I recap the literature on teacher value-added and the influence of education on non-cognitive skills. In Section 1.2 I develop a model in which student outcomes like test scores or attendance are a function of teacher effects, covariates, and random shocks. In Section 1.3 I describe the data and provide descriptive evidence on the pervasiveness of poor attendance in the New York City public schools and correlates of poor attendance. Section 1.4 describes the estimation procedure and the conditions under which parameters of interest are identified. Section 1.5 contains results on the distribution of teacher effects on student attendance and test scores; the persistence of these effects; the usefulness of contemporaneous effects for predicting teacher effects on future student achievement; and factor analysis of the correlation matrix of teacher effects. Section 1.6 demonstrates that results are robust to using a different estimator –
maximum likelihood rather than moment-matching – and to different choices of covariates. Section 2.5 concludes.

1.1 Literature

Skills that aren’t well captured by traditional educational metrics, often termed “non-cognitive skills” or “character skills”, are correlated with many outcomes, including earnings, educational attainment, health, and crime.

However, there is little direct evidence on the degree to which education — and teachers in particular — affects character skills, despite the growing literature on teacher effects on test scores, and the use of test score-based value-added measures in large school systems such as New York City, Los Angeles, Chicago, and Washington, DC. [Jackson (2016) and Gershenson (2016)] examine the effects of teachers on outcomes other than test scores. Gershenson studies third through fifth graders in North Carolina and finds that teachers have “arguably causal, statistically significant effects on student absences that persist over time,” and that "teachers who improve test scores do not necessarily improve student attendance." Similarly, [Jackson (2016)] studies ninth graders in North Carolina and finds that teachers have medium-term effects on student absences, suspensions, grades, and on-time grade progression, and that teacher effects on test scores have modest correlations with teacher effects on behavioral variables. This paper is similar in spirit, but tracks teacher effects over a longer time period.

There are reasons to suspect that teachers may affect their students in the long term in ways that are not captured in test scores. First, increases in the quality or quantity of education are correlated with measures of socioeconomic success, even after conditioning on test scores. For example, [Heckman and Rubinstein (2001)] shows that conditional on AFQT scores, GED recipients earn less than high school graduates who do not attend college. (Chetty et al. 2014b) show that teachers who improve test scores also cause their students to have higher incomes, attend college, and live in better neighborhoods. The mechanisms for this are unclear, since teacher effects on test scores fade out dramatically
after several years (Chetty et al. (2014b), Chetty et al. (2011)); test scores effects alone do not seem sufficient to explain the magnitude of teacher effects on long-term outcomes. Evidence from Project STAR suggests that kindergarten “class quality has significant impacts on non-cognitive measures in fourth and eighth grade such as effort, initiative, and lack of disruptive behavior,” and that high-quality kindergarten classes improve test scores in the short run, but it is not clear whether these effects are due to teacher quality, peer effects, or some other factor (Chetty et al., 2011). In summary, teachers may impact their students’ persistence and motivation, and these effects may be more meaningful or persistent than teacher effects on test scores.

Ultimately, we would like to know how important teachers are as a determinant of meaningful long-term outcomes like high school graduation. Even if we could perfectly measure teacher effects on test scores and other short-term outcomes, these teacher effects are only useful insofar as the proxy for more welfare-relevant effects. In this project, I demonstrate that teachers have significant effects on the medium-term outcomes of future test scores and high school graduation. Short-term teacher effects, however well-measured, are a poor proxy for long-term teacher effects, but teacher effects on attendance and on English test scores do help predict teacher effects on graduation.

1.2 Model

I follow Chamberlain (2013) in developing a model that defines teacher effects as best linear predictor coefficients. In order to preserve independence across teachers, I ensure that the same student does not appear in the data more than once by treating each grade separately and by dropping students who repeat a grade. 3 Thereafter, there is only one observation per student, so I index by student $i$ and use $j(i)$ to refer to the teacher of student $i$ and $c(i)$ to refer to teacher-year cohorts. That is, if $j(i) = j(i')$ but $c(i) \neq c(i')$, then students $i$ and $i'$ are taught by the same teacher but in different years. Subscripts index

---

3 After estimating the covariance of teacher effects for each grade, I pool across grades with a frequency-weighted mean.
observations, and superscripts index outcomes. There are $H = 31$ outcomes: Math test scores, reading test scores, attendance z-scores, four years of leads and four years of lags for each of those variables, and indicators for four-year high school graduation, four-year high school graduation with a Regents diploma, and four-year high school graduation with Advanced Regents diploma. Outcomes are $y_i \in \mathbb{R}^H$ and $x_i \in \mathbb{R}^K$. Teacher $j$’s effect on outcomes, $\mu_j \in \mathbb{R}^H$, is defined as a best linear predictor coefficient. For each outcome $h$,

$$y_i^h = x_i^T \beta^h + \mu^h_{j(i)} + v_i^h$$

In order to interpret parameter estimates causally, we need sorting of students to teachers to be based on observables. In order to estimate $\beta^h$ consistently, we need that errors be orthogonal to covariates and teacher quality: $v_i^h \perp x_i, \mu_{j(i)}$ $\forall h$. In addition, unobservable shocks to outcomes need to be independent of the teacher’s identity, conditional on covariates. This restriction is necessary so that, when estimating variances, we don’t mistake the tendency for some teachers to consistently receive better or worse students for the presence of teachers who consistently teach well or poorly. Imagine that all teachers are identical — $\mu_j = 0$ $\forall j$ — but some teachers are consistently assigned students with high or low shocks. Some teachers will consistently have students who over- or under-perform what would be expected from their covariates, making it appear that teachers vary in quality when they do not.

I further assume that errors across different classrooms are orthogonal: $\mathbb{E} [v_i v_{i'} | c(i) \neq c(i')] = 0$. The parameter of interest is the covariance matrix of teacher effects,

$$\text{Var} (\mu_j) = \Sigma_\mu \in \mathbb{R}^{H \times H}$$ (1.1)

### 1.3 Data, Setting, and Descriptive Statistics

The data includes almost all New York City public school students in grades preschool through 12 in the 2001-02 to 2015-16 school years. I observe rich individual-level data and can track students across years. For each student, I can observe several outcomes of interest:
Test scores, attendance, and what type of diploma the student received.

While my data includes 9.4 million student-year observations that include demographic and attendance information, the effective size of the data is smaller. I can only link teachers to students starting in the 2005-06 school year, although I use observations from earlier years to construct lagged variables and for pre-trend tests. I avoid dropping observations due to missing data in independent variables. Instead, I impute the missing field as the average for that student’s grade and year and also use an indicator variable for missingness. This happens most often when lagged variables are missing because the student recently moved into the district. However, I do require one lagged test score to be present.

I observe demographic information on each student. My data contains each student’s ethnicity and date of birth, which are filled in by parents when the student enters the school system. Other information is recorded by the school administration: grade level, number of days absent, number of days present, and whether the student is in special education. I also observe the student’s registrar data, which explains whether the student is still enrolled, whether the teacher has graduated and with what type of diploma, whether the student has a disability requiring an Individualized Education Program (IEP), and whether the student has dropped out.

New York State has a tiered system of high school diplomas. High school students must take standardized examinations known as Regents exams, and those who pass exams in global history, U.S. history, ELA, math, and science graduate with a “Regents diploma.” Until the 2011-2012 school year, students who met their high school’s graduation but did not meet the requirements for a Regents diploma earned a less prestigious “local diploma”; now, students cannot earn a local diploma unless they have a disability. There also exist diplomas that are harder to attain than a Regents diploma. \(^4\) Students who pass additional exams earn a Regents Diploma with Advanced Designation, and students who satisfy the requirements of the Advanced Designation and attain high scores can attain a “Regents Diploma.”

\(^4\)When the following results refer to a Regents Diploma, this refers to a Regents Diploma that is not with an Advanced Designation, so it is ambiguous whether increasing the number of Regents Diplomas is positive.
with Advanced Designation with Honors."

In addition to the high school Regents exams, students take standardized math and English Language Arts (ELA) tests every year in third through eighth grade. Starting in spring 2013, these tests have been based on Common Core standards.

Math and ELA scores are scored on a scale that varies every year. Therefore, I normalize test scores to have a mean of zero and variance of 1 within each grade and year. Normalization obscures the fact that New York City’s test scores rose dramatically over this period, both in terms of the percentage of students scoring at the proficient level and in comparison to the rest of New York State. Four-year graduation rates rose from 45.5% for students starting ninth grade in 2001 to 70.5% for students starting ninth grade in 2011.

1.3.1 Descriptive Statistics

Table 1.1 contains summary statistics for the whole sample of 10 million students, and Table 1.2 contains summary statistics for the sample that is used for estimation, containing 2.5
Table 1.2: Student summary statistics, for students who can be matched to teachers.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>St. Dev</th>
<th>Min</th>
<th>Max</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>5.98</td>
<td>1.42</td>
<td>4</td>
<td>8</td>
<td>0%</td>
</tr>
<tr>
<td>Year</td>
<td>2009.21</td>
<td>2.08</td>
<td>2006</td>
<td>2013</td>
<td>0%</td>
</tr>
<tr>
<td>Disabled</td>
<td>0.16</td>
<td>0.37</td>
<td>0</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>Female</td>
<td>0.49</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>English Language Learner</td>
<td>0.12</td>
<td>0.32</td>
<td>0</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>Free Lunch</td>
<td>0.83</td>
<td>0.32</td>
<td>0</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>Days absent</td>
<td>11.43</td>
<td>11.95</td>
<td>0</td>
<td>182</td>
<td>0%</td>
</tr>
<tr>
<td>Days present</td>
<td>169.83</td>
<td>12.96</td>
<td>2</td>
<td>186</td>
<td>0%</td>
</tr>
<tr>
<td>Days Absent Lag (Z-Score)</td>
<td>0.06</td>
<td>0.89</td>
<td>-14.66</td>
<td>1.34</td>
<td>2.91%</td>
</tr>
<tr>
<td>Math Score (Z-Score)</td>
<td>0.07</td>
<td>0.97</td>
<td>-6.35</td>
<td>3.89</td>
<td>0%</td>
</tr>
<tr>
<td>Math score lag (Z-Score)</td>
<td>0.06</td>
<td>0.97</td>
<td>-10.00</td>
<td>3.96</td>
<td>5.34%</td>
</tr>
<tr>
<td>ELA Score (Z-Score)</td>
<td>0.05</td>
<td>0.98</td>
<td>-11.10</td>
<td>7.76</td>
<td>0%</td>
</tr>
<tr>
<td>ELA Score Lag (Z-Score)</td>
<td>0.05</td>
<td>0.97</td>
<td>-11.10</td>
<td>7.76</td>
<td>8.08%</td>
</tr>
<tr>
<td>4-Year Graduation</td>
<td>0.68</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
<td>66.47%</td>
</tr>
<tr>
<td>4-Year Graduation, Regents Diploma</td>
<td>0.45</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
<td>66.46%</td>
</tr>
<tr>
<td>4-Year Graduation, Advanced Regents Diploma</td>
<td>0.20</td>
<td>0.40</td>
<td>0</td>
<td>1</td>
<td>66.45%</td>
</tr>
<tr>
<td>N = 2,455,257</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

million students in grades 4 through 8 who can be matched to a teacher. The district is relatively poor, with 77% of students qualifying for free or reduced-price lunch. A plurality of students are Hispanic. Although about 43% of students, as of 2013, speak a language other than English at home, only 13% of students are classified as English Language Learners (ELLs). English Language Learners are students who either take a class in English as a New Language or participate in bilingual education.

Attendance

Student absence is frequent in the New York City schools. Descriptive, non-causal evidence suggests that attendance matters for student achievement. Poor attendance is associated with lower test scores and a lower likelihood of graduating high school, and there are large socioeconomic gaps in attendance. Although this data cannot explain why students miss school so often, it is consistent with the hypothesis that students are absent far more often than necessitated by illness. The average New York City public school student is absent 16
days in an approximately 180-day school year, or 9%. By comparison, the average student nationally is absent on about 7% of days.

(a) Empirical CDF of absences in elementary schools (grades PK through 4), middle schools (grades 5 through 8), and high schools (grades 9-12).

(b) Percentiles of attendance in each grade.

Figure 1.1

Attendance deteriorates dramatically across grades, as shown in Figure 1.1a. In each grade, about 8% of students are never absent. However, high school students are absent far more often than elementary or middle school students, especially in the right tail. Figure 1.1b shows this by plotting the fifth, fiftieth, and ninety-fifth percentiles of absences within each grade.

Figure 1.2 illustrates the large ethnic gaps in school attendance. In any grade, Hispanic, Black, and Native American students are absent almost twice as often as Asian students, with non-Hispanic White and multi-racial students in the middle.

Tables 1.3, 1.4, 1.5, and 1.6 show coefficients from regressing ELA scores, math scores, attendance, and high school graduation on student demographic characteristics, with and without lagged values of outcomes. All outcomes have been z-scored within grade-year, so coefficients are comparable in magnitude. A student with an attendance z-score of 1 is present one standard deviation more than other students in her grade and year. Although coefficients on regressions with ELA scores (Table 1.3) and with math scores (Table 1.4) as the dependent variable have similar coefficients, they are not similar to the coefficients
Figure 1.2: *Average number of days absent by ethnicity.*
Table 1.3: Predictors of English Language Arts scores.

Notes. For categorical variables – ethnicity, home language, and year – an F-statistic and its p-value are shown. Other variables have standard errors in parentheses. *** indicates significance at the 0.1% level.
### Table 1.4: Predictors of math scores.

*Notes.* For categorical variables – ethnicity, home language, and year – an F-statistic and its p-value are shown. Other variables have standard errors in parentheses. *** indicates significance at the 0.1% level.
<table>
<thead>
<tr>
<th></th>
<th>Attendance (Normalized)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Constant</td>
<td>0.462***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
</tr>
<tr>
<td>English Language Learner</td>
<td>-0.046***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Female</td>
<td>0.027***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Free Lunch</td>
<td>-0.125***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>IEP</td>
<td>-0.306***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Lag ELA Score</td>
<td>-0.003***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Lag Math Score</td>
<td>-0.01***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Lagged Attendance</td>
<td>0.645***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Lagged Values Missing</td>
<td>-0.095***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>18691.9</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
</tr>
<tr>
<td>Home Language</td>
<td>11216.6</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
</tr>
<tr>
<td>Year</td>
<td>26.3365</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
</tr>
<tr>
<td>N</td>
<td>9.4M</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.064287</td>
</tr>
</tbody>
</table>

**Table 1.5: Predictors of attendance.**

*Notes.* For categorical variables – ethnicity, home language, and year – an F-statistic and its p-value are shown. Other variables have standard errors in parentheses. *** indicates significance at the 0.1% level.
<table>
<thead>
<tr>
<th></th>
<th>Graduation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.513*</td>
<td>0.355*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.31)</td>
<td>(0.284)</td>
<td></td>
</tr>
<tr>
<td>English Language Learner</td>
<td>-0.156***</td>
<td>-0.032***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.07***</td>
<td>0.058***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td></td>
</tr>
<tr>
<td>Free Lunch</td>
<td>-0.097***</td>
<td>-0.034***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td></td>
</tr>
<tr>
<td>IEP</td>
<td>-0.206***</td>
<td>-0.032***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td></td>
</tr>
<tr>
<td>Attendance</td>
<td>0.143***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronically Absent</td>
<td>0.003*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELA score</td>
<td>0.059***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math score</td>
<td>0.107***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>2441.58</td>
<td>343.839</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td></td>
</tr>
<tr>
<td>Home Language</td>
<td>328.433</td>
<td>55.3449</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>213.836</td>
<td>198.975</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>823370</td>
<td>823370</td>
<td></td>
</tr>
<tr>
<td>R-squared</td>
<td>0.112862</td>
<td>0.255901</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.6: Predictors of graduation.

Notes. For categorical variables – ethnicity, home language, and year – an F-statistic and its p-value are shown. Other variables have standard errors in parentheses. *** indicates significance at the 0.1% level.
from predicting attendance (Table 1.5). The only coefficients that are large in magnitude for predicting attendance are coefficients on lagged values of outcomes and having an Individualized Educational Plan (an IEP, for students with disabilities). Despite the large ethnic gaps in school attendance, attendance is hard to predict (within grade-years), with an $R^2$ of only 0.06 in a regression that includes ethnicity, indicators for common home languages, year dummies, and a student’s status as disabled, ELL, receiving free lunch, female, or in special education. $R^2$ rises to 0.179 in a regression that includes lagged values of attendance and test scores. Test scores, however, are much more predictable, generating $R^2$ values of 0.30 to 0.49 across specifications.

Table 1.6 shows that attendance and test scores are helpful in predicting graduation. Attendance is a better predictor than test scores: a one standard deviation increase in
attendance corresponds to a 14 percentage point increase in graduation rates, while one standard deviation increases in math or ELA test scores correspond to 11 percentage point and 6 percentage point increases, respectively.

Figure 1.3 displays the relationship between graduation, absences, math scores, and ELA scores nonparametrically. The top left panel is a binned scatter plot of graduation against absences, with one point for each integer number of absences. The middle and bottom left panels repeat this with math and English Language Arts test scores as the y variable. The right-side plots are recentered residual plots, controlling for English Language Learner status, gender, free lunch status, whether the student has an Individualized Education Plan, ethnicity, home language, year, and grade level. The figure plots residuals of the outcome plus the mean outcome against residuals attendance plus mean attendance. Technically, define \( \tilde{\text{graduation}}_i \) and \( \tilde{\text{absences}}_i \), where \( \tilde{\text{graduation}}_i = \text{graduation}_i - E^*[\text{graduation}_i|1, z_i] \), \( E^* \) is the best linear predictor operator, and \( z \) includes control variables. The regression line plots

\[
E^* [\text{graduation}_i|1, \text{absences}_i = \tilde{\text{absences}}_i + E^* [\text{absences}_i|z_i = \bar{z}]].
\]

against \( \tilde{\text{absences}}_i + E^* [\text{absences}_i|z_i = \bar{z}] \). The right panel answers the question, "what would the left panel look like if everyone had the same covariates?"

Figure 1.3 suggests a strong relationship between attendance and graduation, math, and ELA scores, at least for students absent less than fifty days per year. For students absent more than fifty days per year, the marginal cost of poor attendance appears to decline – perhaps because these students are already doing quite poorly, with a less than 10% chance of graduating high school and test scores nearly one standard deviation below the mean. On average, ten more days of attendance corresponds to an 8 percentage point greater chance of graduating high school (7 percentage points with controls), a gain of 0.25 standard deviations in math test scores (0.17 with controls), and 0.19 standard deviations in ELA test scores (0.11 with controls). These estimates are clearly not causal, but are perhaps an upper bound on the gains in graduation and test scores that might be expected to come from improved attendance.
1.4 Estimation

I estimate the covariance of teacher effects, $\Sigma_\mu$, using a moment-matching procedure similar to that of Kane and Staiger (2008). In order to preserve independence across teachers, I prevent the same student from appearing in the data more than once by running estimation separately for each grade and by dropping students for repeating a grade. The first step in estimation is residualizing outcomes by estimating the $\beta$ of Equation ?? by regressing within-teacher variation in outcomes on within-teacher variation in covariates:

$$\hat{\beta}^h = \arg\min_\beta \sum_i \left( y_i^h - \bar{y}_{j(i)}^h - \left( x_i - \bar{x}_{j(i)} \right)^T \beta \right)^2$$

Key to estimating $\Sigma_\mu$ is that the errors in Equation ?? are independent across different classrooms, so the average product of residuals in classrooms taught by the same teacher is the covariance of teacher effects. Let $C(j)$ be the set of all classrooms taught by teacher $j$, and let $\bar{y}_c$ and $\bar{x}_c$ be average outcomes and covariates in classroom $c$. When errors are independent across classrooms, then

$$E \left[ \left( \bar{y}_c - \bar{x}_c^T \hat{\beta} \right) \left( \bar{y}_{c'} - \bar{x}_{c'}^T \hat{\beta} \right)^T \mid c, c' \in C(j), c \neq c' \right] = \Sigma_\mu.$$  

Using a “moment-matching” procedure as in Kane and Staiger (2008) and Chamberlain (2013), $\hat{\Sigma}_\mu$ is the average variance of residualized outcomes in unique pairs of different classrooms taught by the same teacher:

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

1.4.1 Inference

I estimate the posterior distribution of $\Sigma_\mu$ using the Bayesian Bootstrap (Rubin, 1981). In the $n$th Bayesian Bootstrap draw, reweight the data with teacher-level weights $\omega^n \in \mathbb{R}^N$ drawn $\omega^n \sim \text{Dirichlet}(1, 1, \ldots, 1)$. First, estimate

$$\hat{\beta}^{h,n} = \arg\min_\beta \sum_i \omega_{j(i)}^n \left( y_i - \bar{y}_{j(i)} - \left( x_i - \bar{x}_{j(i)} \right)^T \beta \right)^2$$
Then the $n^{th}$ draw of $\Sigma_\mu$ is

$$\hat{\Sigma}_n = \frac{2}{\sum_j \omega_j^n \mid C(j) \mid (\mid C(j) \mid - 1)} \sum_j \omega_j^n \sum_{c,c' \in C(j); c \neq c'} \left( g_c - \bar{x}_c^T \hat{\beta} \right) \left( g_{c'} - \bar{x}_{c'}^T \hat{\beta} \right)^T$$

### 1.4.2 Identification

In order to understand how teachers contribute to variation in student outcomes, we must ensure that teachers receive credit or blame only for changes in outcomes they cause, and not for changes that would have happened with an average teacher. A threat to our ability to identify the variance of teacher value-added is systematic sorting of students to teachers.

This “sorting on observables” restriction, discussed formally in Section 1.2, requires that random shocks be independent of a teacher’s identity, conditional on covariates. I include a rich set of controls to make the sorting on observables restriction plausible. I estimate the model separately for each grade level, so coefficients can change from year to year to allow for, for instance, the persistence of test scores to vary with grade. I control for age, gender, limited English proficiency status, free or reduced-price lunch eligibility, twice-lagged absences, twice-lagged test scores, disability status, teacher-grade level averages of all of these variables, indicators for ethnicity, indicators for year, and indicators for the ten most common home languages. Rather than drop students with missing data, I set missing variables to zero and control for indicators for missingness. I do, however, drop students missing lagged attendance or test scores.

My controls are similar to those in Chetty et al. (2014a). There are three differences between their controls and mine: They control for suspensions, which I do not observe; they estimate their model on all grades and interact many variables with grade dummies; and they control for cubics in previous test scores and classroom means of previous test scores.

To ensure that my results are robust to using their controls, I generate point estimates of the diagonal of $\Sigma_\mu$ while additionally controlling for cubics in previous test scores, previous attendance, and classroom means of previous test scores and previous attendance. These results are in Section 1.6 controlling for cubics does not substantially change results.

Although the sorting on observables restriction cannot be directly tested, I follow Chetty
et al. (2014a) in using pre-trend tests to test whether students show unusual improvement or declines in outcomes in the years before being assigned to a higher value-added teacher. Pre-trend coefficients, discussed below, correspond to negative years in Figures 1.5 and A.2 and in Tables 1.9 and A.3. Visually, pre-trends appear small, but they are statistically significant, with zero not lying in the 95% credible set. These results suggest that controlling for more lags could (but might not) mitigate sorting on observables. In Section 1.6 I estimate the diagonal of $\Sigma_\mu$ while controlling for all available lags in addition to the baseline controls; additional lags to not appear to affect the results at all.

### 1.5 Results

I conducted analysis separately for math teachers and English teachers. Results for math and English teachers currently look very similar, because the sample includes many fourth and fifth grade teachers who teach both math and English. Figures for math teachers are in this section and figures for English teachers are in Appendix A.

The square root of the diagonal of $\Sigma_\mu$ is the standard deviation of teacher effects on each outcome. For example, the square root of the diagonal element of $\Sigma_\mu$ corresponding to four-year high school graduation rates is 0.049, indicating that students of a teacher who is one standard deviation above average at improving graduation rates are 4.9 percentage points higher.

<table>
<thead>
<tr>
<th>Year</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA score</td>
<td>0.112</td>
<td>0.103</td>
<td>0.117</td>
<td>0.117</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>(0.108, 0.117)</td>
<td>(0.099, 0.108)</td>
<td>(0.112, 0.124)</td>
<td>(0.109, 0.125)</td>
<td>(0.080, 0.121)</td>
</tr>
<tr>
<td>Math score</td>
<td>0.174</td>
<td>0.134</td>
<td>0.144</td>
<td>0.141</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td>(0.170, 0.179)</td>
<td>(0.131, 0.139)</td>
<td>(0.139, 0.150)</td>
<td>(0.132, 0.149)</td>
<td>(0.121, 0.161)</td>
</tr>
<tr>
<td>Attendance Z-Score</td>
<td>0.069</td>
<td>0.070</td>
<td>0.079</td>
<td>0.080</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>(0.067, 0.073)</td>
<td>(0.067, 0.073)</td>
<td>(0.075, 0.083)</td>
<td>(0.075, 0.086)</td>
<td>(0.061, 0.088)</td>
</tr>
</tbody>
</table>

**Table 1.7: Magnitude of math teacher effects on test scores and attendance.**

**Notes:** The standard deviation of math teacher effects on test scores and attendance, one to four years out. This table displays the same information as Figure 1.4. 95% credible interval from 1000 Bayesian Bootstrap iterations in parentheses.
Figure 1.4: Magnitude of math teacher effects on test scores and attendance.

Notes: The top plot shows the standard deviation of math teachers' effects on math scores, in the same year that the student has this teacher and 1, 2, 3, and 4 years after. The second and third plots show the variances of math teachers’ effects on English Language Arts scores and attendance. Error bars plot a 95% credible interval based on 1000 Bayesian Bootstrap iterations.

<table>
<thead>
<tr>
<th>Standard Deviation of Teacher Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduated, 4-year</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Regents Diploma, 4-year</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Advanced Regents Diploma, 4-year</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 1.8: Magnitude of math teacher effects on high school graduation.

Notes: The standard deviation of math teacher effects on four-year high school graduation. 95% credible interval in parentheses.
points more likely to graduate high school. The variance of teacher effects on high school graduation is $0.0024 (0.049^2)$, accounting for 1% of the variance in high school graduation rates. Figure 1.4, Table 1.7, and Table 1.8 show the effects of math teachers on math scores, ELA scores, and attendance. The standard deviation of math teacher effects on contemporaneous math scores is about 0.174, roughly in line with previous results, indicating that math teachers account for 3% of variance in same-year math test scores ($0.174^2$). The standard deviation of math teacher effects is slightly smaller in all succeeding years, around 0.14. By contrast, teacher effects on English Language Arts test scores are smaller – Appendix Table A.1 shows the analog of 1.4 for ELA teachers – but are more consistent across years. One explanation for these different patterns in math and ELA scores is that the subject matter on math tests overlaps less across grades than the subject matter on ELA tests. For example, grades 6 through 8 share the same Common Core ELA standards, which cover broad topics like “Determine the central ideas or information of a primary or secondary source,” while the math topics change every year and cover narrower topics like “rational and irrational numbers.”

For attendance, the standard deviation of math teacher effects is about 0.07 in each year. Attendance, like test scores, has been Z-scored within each grade and year, so this implies that having a teacher who is one standard deviation above average at increasing attendance increases students’ attendance by 0.07 of a standard deviation relative to their peers.

We can also ask, given a teacher’s effect on outcome $h'$, what is her expected effect on outcome $h$? Other work has addressed this question by regressing outcomes on estimated teacher effects and controls, but it can also be answered using only $\Sigma_\mu$, if we update Equation 1.1 to assume homoskedasticity:

$$\mathbb{E} \left[ H_{j(i)} H_{j(i)}^T | x_i \right] = \Sigma_\mu.$$  

(1.2)
Figure 1.5: Fade out of math teacher effects.

Notes: The best linear predictor coefficient for predicting a teacher’s effect on a future outcome given her effect on a present outcome and covariates. Error bars plot a 95% credible interval based on 1000 Bayesian Bootstrap iterations.
Table 1.9: Fade out of math teacher effects.

<table>
<thead>
<tr>
<th>Year</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA</td>
<td>0.139</td>
<td>0.155</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>(0.063, 0.219)</td>
<td>(0.088, 0.217)</td>
<td>(0.000, 0.000)</td>
<td>(0.000, 0.000)</td>
<td>(1.000, 1.000)</td>
</tr>
<tr>
<td>Math</td>
<td>-0.067</td>
<td>-0.023</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>(-0.106, -0.026)</td>
<td>(-0.051, 0.006)</td>
<td>(0.000, 0.000)</td>
<td>(0.000, 0.000)</td>
<td>(1.000, 1.000)</td>
</tr>
<tr>
<td>Attendance Z-Score</td>
<td>0.106</td>
<td>0.094</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>(0.031, 0.184)</td>
<td>(0.044, 0.145)</td>
<td>(0.000, 0.000)</td>
<td>(0.000, 0.000)</td>
<td>(1.000, 1.000)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Graduated (4-year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA</td>
<td>0.532</td>
<td>0.451</td>
<td>0.294</td>
<td>0.347</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>(0.470, 0.593)</td>
<td>(0.372, 0.532)</td>
<td>(0.201, 0.386)</td>
<td>(0.186, 0.510)</td>
<td>(-0.060, 0.108)</td>
</tr>
<tr>
<td>Math</td>
<td>0.360</td>
<td>0.161</td>
<td>0.049</td>
<td>-0.064</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.333, 0.389)</td>
<td>(0.119, 0.199)</td>
<td>(-0.001, 0.095)</td>
<td>(-0.150, 0.022)</td>
<td>(-0.038, 0.024)</td>
</tr>
<tr>
<td>Attendance Z-Score</td>
<td>0.455</td>
<td>0.313</td>
<td>0.165</td>
<td>0.111</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>(0.400, 0.510)</td>
<td>(0.245, 0.380)</td>
<td>(0.095, 0.239)</td>
<td>(0.009, 0.211)</td>
<td>(-0.001, 0.205)</td>
</tr>
</tbody>
</table>

Notes: Best linear predictor coefficients. 95% credible interval based on 1000 Bayesian Bootstrap iterations in parentheses. Coefficients on outcomes from years -2 and -1 are zero because lags are controlled for, and the coefficient on the year 0 outcome is 1 by construction.

Combining Equations ?? and [1.2]

\[
E^* \left[ y_{ij}^h | \mu_{ij}^{h'}, x_i \right] = E^* \left[ \mu_{ij}^H | \mu_{ij}^{h'}, x_i \right] + x_i^T \beta^h
\]

\[
= \frac{\text{Cov} \left( \mu_j \right)_{h,h'} \mu_{ij}^{h'} + x_i^T \beta^h}{\text{Var} \left( \mu_j^{h'} \right)}
\]

\[
= \frac{\Sigma_{\mu_{ij}^{h'}} \mu_{ij}^{h'} + x_i^T \beta^h}{\Sigma_{\mu_{ij}^{h'}} \mu_{ij}^{h'} + x_i^T \beta^h}
\]

\[
\equiv \gamma_{h,h'} \mu_j^{h'} + x_i^T \beta^h
\]

Figure 1.5 plots $\gamma_{h,h'}$ where $h =$ contemporaneous English Language Arts scores, math scores, and attendance and $h' =$ represents leads and lags of those outcomes (again using only math teachers). Since controls include lagged test scores and attendance, the coefficients on previous-year value-added is zero. Figure 1.5 and Table 1.9 show that math teachers who improve math scores by $x$ are expected to improve their students’ scores in the next year by less than half of $x$, and their scores in the year after by very little. This is a similar result to that
found in [Chetty et al. (2014a)]. Combined with Figure 1.4 we see that although teachers do vary in their effects on their students’ test scores four years in the future, very little of that variation is captured by teacher effects on same-year test scores. That is, there are teachers who are much better or worse than average at boosting long-term test scores, but they not especially likely to be the teachers who raise short-term test scores. There appears to be more fade-out for math scores than for attendance and more for attendance than for ELA scores. The faster fade-out for math scores is perhaps not surprising, since the topics tested on Common Core ELA tests are more similar between grades for ELA than for math. Figure A.2 and Table A.3 repeat the same figure and table for English teachers.

Does $\Sigma_\mu$ vary by grade?

**Figure 1.6: Magnitude of math teacher effects by grade.**

Notes: The standard deviation of math teachers’ effects on outcomes, or the square root of the diagonal of $\Sigma_\mu$, within each year. Error bars plot a 95% credible interval based on 1000 Bayesian Bootstrap iterations.
Figure 1.6 and Appendix Figure A.3 show the variance of teacher effects on each outcome, like Figure 1.4, but broken down by grade. It appears that teachers who teach later grades have smaller effects on ELA scores, but there is no consistent pattern for math scores or attendance. This is consistent with the hypothesis that ELA tests are more cumulative than math tests.

Figure 1.7: Fade out of math teacher effects, by grade.

Covariance of time-zero VA with future VA

Notes: The best linear predictor coefficient for predicting a teacher’s effect on a future outcome given her effect on a present outcome and covariates, by grade. Error bars plot a 95% credible interval based on 1000 Bayesian Bootstrap iterations.

Figure 1.7 and Appendix Figure A.4 show the best linear predictor coefficients for predicting a teacher’s effect on a future or past outcome given her effect on current outcomes. Generally, teachers of later grades appear to have more persistent effects on test scores, and there is no obvious pattern for attendance. These results are in contrast with Heckman (2012)’s thesis that cognitive abilities are relatively more malleable than non-cognitive...
abilities at younger ages, which would have predicted smaller or less persistent effects on test scores in later grades, but the opposite for attendance.

1.5.1 Are current teacher effects a good proxy for future teacher effects?

How well can teacher effects on a long-term outcome be captured by teacher effects on short-term outcomes? We can answer this question by constructing an $R^2$-like statistic, “goodness of proxy,” that reflects how much of the variation in some long-term teacher effect $h$ is captured by short-term teacher effects on outcomes in set $Q$. If we assume that $\mu_j$ has a multivariate normal distribution, then the expectation of one component of $\mu_j$ given other components is linear and can be expressed as a function of the covariances in $\Sigma_{\mu}$. Say we are interested in predicting value-added at component $h$, $\mu^h_j$, and know a vector-valued value-added $\mu^Q_j$ for outcomes in set $Q$. Then the expectation of $\mu^h_j$ given $\mu^Q_j$ is

$$E[\mu^h_j|\mu^Q_j] = E^*[\mu^h_j|\mu^Q_j] = (\mu^Q_j)^T \text{Var}(\mu^Q)^{-1} \text{Cov}(\mu^Q, \mu^h)$$

Equation 1.3 defines a “goodness of proxy” statistic that is zero when $\mu^Q_j$ does not help predict $\mu^h_j$ and equals one when it is perfectly predictable. This measure is also computable using only the components of $\Sigma_{\mu}$.

$$\text{goodness of proxy}_{h,Q} \equiv 1 - \frac{E[\text{Var}(\mu^h_j|\mu^Q_j)]}{\text{Var}(\mu^h)} = \frac{\text{Var}(E[\mu^h_j|\mu^Q_j])}{\text{Var}(\mu^h)} = \frac{\text{Cov}(\mu^Q, \mu^h)^T \text{Var}(\mu^Q)^{-1} \text{Cov}(\mu^Q, \mu^h)}{\text{Var}(\mu^h)}$$

Equation 1.3 defines a “goodness of proxy” statistic that is zero when $\mu^Q_j$ does not help predict $\mu^h_j$ and equals one when it is perfectly predictable. This measure is also computable using only the components of $\Sigma_{\mu}$.

Table 1.10 and Appendix Table A.4 show the goodness of proxy for graduation and four-year-lead test scores and attendance, using same-year test scores and attendance as a proxy. Teacher effects on four-year-ahead ELA scores are somewhat predictable, with a Goodness of
Table 1.10: Goodness of proxy.

<table>
<thead>
<tr>
<th></th>
<th>(1) Scores</th>
<th>(2) Attendance</th>
<th>(3) Scores + Attend</th>
<th>(4) Ratio (3) / (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA score (4 years later)</td>
<td>0.244</td>
<td>0.072</td>
<td>0.294</td>
<td>1.205</td>
</tr>
<tr>
<td></td>
<td>(0.107, 0.459)</td>
<td>(0.008, 0.191)</td>
<td>(0.140, 0.525)</td>
<td>(1.008, 1.867)</td>
</tr>
<tr>
<td>Math score (4 years later)</td>
<td>0.037</td>
<td>0.006</td>
<td>0.043</td>
<td>1.146</td>
</tr>
<tr>
<td></td>
<td>(0.009, 0.104)</td>
<td>(0.000, 0.049)</td>
<td>(0.013, 0.125)</td>
<td>(1.001, 2.755)</td>
</tr>
<tr>
<td>Attendance Z-Score (4 years later)</td>
<td>0.028</td>
<td>0.011</td>
<td>0.037</td>
<td>1.311</td>
</tr>
<tr>
<td></td>
<td>(0.004, 0.079)</td>
<td>(0.000, 0.041)</td>
<td>(0.010, 0.093)</td>
<td>(1.004, 4.827)</td>
</tr>
<tr>
<td>Graduated, 4-year</td>
<td>0.013</td>
<td>0.021</td>
<td>0.033</td>
<td>2.471</td>
</tr>
<tr>
<td></td>
<td>(0.001, 0.072)</td>
<td>(0.000, 0.072)</td>
<td>(0.006, 0.118)</td>
<td>(1.036, 28.489)</td>
</tr>
<tr>
<td>Regents Diploma, 4-year</td>
<td>0.014</td>
<td>0.000</td>
<td>0.015</td>
<td>1.058</td>
</tr>
<tr>
<td></td>
<td>(0.001, 0.076)</td>
<td>(0.000, 0.014)</td>
<td>(0.002, 0.077)</td>
<td>(1.000, 8.068)</td>
</tr>
<tr>
<td>Advanced Regents Diploma, 4-year</td>
<td>0.059</td>
<td>0.006</td>
<td>0.062</td>
<td>1.040</td>
</tr>
<tr>
<td></td>
<td>(0.029, 0.096)</td>
<td>(0.000, 0.019)</td>
<td>(0.031, 0.101)</td>
<td>(1.000, 1.217)</td>
</tr>
</tbody>
</table>

Notes: Goodness of proxy for graduation and four-year-lead test scores and attendance, using same-year test scores and attendance. 95% credible set based on 1000 Bayesian Bootstrap iterations in parentheses.

Proxy of 0.294 (95% CI [0.140, 0.525]), but teacher effects on all other longer-term outcomes are much less predictable, with a Goodness of Proxy of 0.015 to 0.062. This is consistent with Tables 1.7 and Appendix Table A.1 which show that teachers vary substantially in their effects on test scores and attendance four years in the future, and with Tables 1.9 and A.3 which show that teachers who improve math test scores or attendance do not have persistent effects, while teachers who improve ELA scores do.

1.5.2 Bivariate Correlations and Factor Analysis

Can $\Sigma_{\mu}$ be represented well by several factors? What are they? I answer this question by transforming the covariance matrix $\Sigma_{\mu}$ matrix to a correlation matrix $\tilde{\Sigma}_{\mu}$, and estimating the factors of $\tilde{\Sigma}_{\mu}$ via maximum likelihood (Jöreskog 1970). I use only the component of the matrix that corresponds to non-negative years, since these are the years in which we expect teachers to have real effects.

Before applying factor analysis, we can gain some intuition for what the factors might look like from Table 1.11 which shows the correlation structure of teacher effects on different present-year outcomes. Although teacher effects on math and reading test scores are highly
Table 1.11: Correlations between teachers’ value-added on different outcomes.

<table>
<thead>
<tr>
<th></th>
<th>ELA score</th>
<th>Math score</th>
<th>Attendance</th>
<th>4-year Grad</th>
<th>4-yr Reg. Dip.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA score</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1.00, 1.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math score</td>
<td>0.66</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.63, 0.68)</td>
<td>(1.00, 1.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attendance</td>
<td>0.15</td>
<td>0.12</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.10, 0.20)</td>
<td>(0.08, 0.17)</td>
<td>(1.00, 1.00)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-year Grad</td>
<td>0.08</td>
<td>-0.01</td>
<td>0.15</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>(-0.12, 0.22)</td>
<td>(-0.12, 0.08)</td>
<td>(0.00, 0.27)</td>
<td>(1.00, 1.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-yr Reg. Dip.</td>
<td>-0.11</td>
<td>-0.05</td>
<td>0.01</td>
<td>0.52</td>
<td>1.00</td>
</tr>
<tr>
<td>(-0.27, 0.00)</td>
<td>(-0.15, 0.02)</td>
<td>(-0.09, 0.11)</td>
<td>(0.46, 0.63)</td>
<td>(1.00, 1.00)</td>
<td></td>
</tr>
<tr>
<td>4-yr Adv. Reg. Dip.</td>
<td>0.22</td>
<td>0.07</td>
<td>0.08</td>
<td>0.15</td>
<td>-0.74</td>
</tr>
<tr>
<td>(0.15, 0.28)</td>
<td>(0.02, 0.12)</td>
<td>(0.01, 0.14)</td>
<td>(0.04, 0.21)</td>
<td>(-0.77, -0.66)</td>
<td></td>
</tr>
</tbody>
</table>

correlated, with a correlation of 0.66, teacher effects on test scores and on attendance are much lower.

This low correlation indicates that my findings but do not resolve the “fade-out mystery” of teacher effects on test scores: if students of high score-VA teachers have only slightly improved test scores four years later, how is it that teachers have substantial effects on later student outcomes? Since teachers who improve test scores are not much better than average at improving attendance, the fade-out puzzle cannot be resolved by high score-VA teachers improving attendance.

In factor analysis there are several rules of thumb for choosing the number of components. One is to count the number of eigenvalues that are greater than one. This heuristic suggests four components for both math and English teachers. Another heuristic suggests plotting the eigenvalues, as in Figures 1.8 and A.5 and look for a point where the eigenvalues start to level off. This heuristic suggests three factors for math teachers and three for ELA teachers. For simplicity of exposition, I choose three factors.

Figure 1.9 shows the estimated factors for math teachers. (Appendix Figure A.6 is the analog for ELA teachers; it is similar.) The first factor is associated with positive effects on all factors except for same-year math scores, and especially large effects on long-term test
Figure 1.8: Eigenvalues of $\tilde{\Sigma}_\mu$ for math teachers.

Notes: Error bars show a 95% credible interval based on 1000 Bayesian Bootstrap iterations.
Factors of math teacher effects

Figure 1.9: Factors of $\Sigma_h$ for math teachers.
scores. A teacher whose effects resemble this component has positive long-term effects on her students – or negative, as the sign of the factor is not identified – but her effectiveness would not be noted by focusing on present-year outcomes. The second factor reflects positive effects on long-term test scores, especially in math, and on graduation, especially with an Advanced Regents designation, but negative effects on attendance and short-term test scores. The third factor has large, same-signed loadings on all test scores, but small loadings on attendance and graduation.

1.6 Robustness Checks

Various methods are available to estimate the diagonal of $\Sigma_t$, which reveals the variance of teacher effects on each outcome, as surveyed in Chapter 2 of this dissertation. And of course, infinitely many sets of control variables are possible. This section estimates the square root of the diagonal of $\Sigma_t$ for two different estimators and three sets of controls for math teachers; results from English teachers are in the appendix.

The method used in this paper is a “moment-matching” estimator that residualizes test scores using within-teacher variation and compares mean residuals in different classrooms taught by the same teacher. This method can be expressed as an instance of the Generalized Method of Moments and is identical to the “modified-KS” estimator in Chapter 2. An alternative estimation method is maximum likelihood. See Figure 1.10 and Appendix Figure A.7.

In the “baseline” controls used to generate the results in Section 1.5, I control for age, gender, limited English proficiency status, free or reduced-price lunch eligibility, twice-lagged absences, twice-lagged test scores, disability status, teacher-grade level averages of all of these variables, indicators for ethnicity, indicators for year, and indicators for the ten most common home languages. Rather than drop students with missing data, I set missing variables to zero and control for indicators for missingness. I do, however, drop students missing lagged attendance or test scores. Figure 1.11 and Appendix Figure A.8 additionally show results from two other sets of control variables. The “Polynomial” controls control for
Figure 1.10: Robustness to choice of estimator.

Notes: Standard deviation of math teacher effects on present and future outcomes, for both the “moment-matching” estimator used above and maximum likelihood.
all baseline variables, cubics in lagged test scores and attendance, and cubics in teacher-year level means of lagged test scores and attendance. The “Polynomial” controls are very similar to the controls used in Chetty et al. (2014a). The “All Lags” controls control for up to four years of lags, with missing lags imputed to zero and an indicator for missingness.

**Standard deviation of math teacher effects: Robustness to Control**

![Graph showing the standard deviation of math teacher effects over time with different sets of controls.](image)

**Figure 1.11: Robustness to choice of controls.**

*Notes:* Standard deviation of math teacher effects on present and future outcomes, for the baseline controls used above, baseline controls plus all available lags, and for the controls from Chetty et al. (2014a), which include third-degree polynomials.

Tables 1.12, 1.13, 1.14, and 1.15 (and Appendix Tables A.5, A.6, A.7, and A.8) display all of the results found in the figures, as well as results from maximum likelihood estimation with the non-baseline sets of controls.

Figure 1.10 and Tables 1.12 through 1.15 show that using the likelihood-based estimator instead of the moment-matching estimator yields similar but uniformly larger results. Figure 1.11 and Tables 1.12 through 1.15 show that different sets of controls generally yield very...
<table>
<thead>
<tr>
<th>Controls</th>
<th>Estimator</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Moment-Matching</td>
<td>0.116 0.106 0.118 0.119 0.106</td>
</tr>
<tr>
<td>Baseline</td>
<td>MLE</td>
<td>0.118 0.109 0.124 0.127 0.126</td>
</tr>
<tr>
<td>Polynomial</td>
<td>Moment-Matching</td>
<td>0.116 0.101 0.111 0.116 0.175</td>
</tr>
<tr>
<td>Polynomial</td>
<td>MLE</td>
<td>0.120 0.113 0.123 0.137 0.197</td>
</tr>
<tr>
<td>All Lags</td>
<td>Moment-Matching</td>
<td>0.115 0.105 0.117 0.118 0.105</td>
</tr>
<tr>
<td>All Lags</td>
<td>MLE</td>
<td>0.130 0.110 0.124 0.128 0.126</td>
</tr>
</tbody>
</table>

**Table 1.12:** Standard deviation of math teacher effects on ELA scores: Robustness to choice of controls and estimator.

<table>
<thead>
<tr>
<th>Controls</th>
<th>Estimator</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Moment-Matching</td>
<td>0.178 0.140 0.152 0.152 0.155</td>
</tr>
<tr>
<td>Baseline</td>
<td>MLE</td>
<td>0.191 0.155 0.165 0.162 0.156</td>
</tr>
<tr>
<td>Polynomial</td>
<td>Moment-Matching</td>
<td>0.186 0.144 0.157 0.171 0.355</td>
</tr>
<tr>
<td>Polynomial</td>
<td>MLE</td>
<td>0.200 0.163 0.177 0.180 0.390</td>
</tr>
<tr>
<td>All Lags</td>
<td>Moment-Matching</td>
<td>0.178 0.140 0.151 0.151 0.154</td>
</tr>
<tr>
<td>All Lags</td>
<td>MLE</td>
<td>0.194 0.155 0.165 0.162 0.159</td>
</tr>
</tbody>
</table>

**Table 1.13:** Standard deviation of math teacher effects on math scores: Robustness to choice of controls and estimator.

<table>
<thead>
<tr>
<th>Controls</th>
<th>Estimator</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Moment-Matching</td>
<td>0.072 0.074 0.081 0.081 0.075</td>
</tr>
<tr>
<td>Baseline</td>
<td>MLE</td>
<td>0.078 0.084 0.090 0.091 0.097</td>
</tr>
<tr>
<td>Polynomial</td>
<td>Moment-Matching</td>
<td>0.077 0.077 0.085 0.094 0.120</td>
</tr>
<tr>
<td>Polynomial</td>
<td>MLE</td>
<td>0.087 0.086 0.092 0.098 0.127</td>
</tr>
<tr>
<td>All Lags</td>
<td>Moment-Matching</td>
<td>0.070 0.072 0.079 0.080 0.074</td>
</tr>
<tr>
<td>All Lags</td>
<td>MLE</td>
<td>0.083 0.082 0.089 0.090 0.101</td>
</tr>
</tbody>
</table>

**Table 1.14:** Standard deviation of math teacher effects on attendance: Robustness to choice of controls and estimator.
Table 1.15: Standard deviation of math teacher effects on graduation: Robustness to choice of controls and estimator.

<table>
<thead>
<tr>
<th>Controls</th>
<th>Estimator</th>
<th>Graduated</th>
<th>Regents Diploma</th>
<th>Advanced Regents Diploma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Moment-Matching</td>
<td>0.049</td>
<td>0.074</td>
<td>0.065</td>
</tr>
<tr>
<td>Baseline</td>
<td>MLE</td>
<td>0.051</td>
<td>0.076</td>
<td>0.065</td>
</tr>
<tr>
<td>Polynomial</td>
<td>Moment-Matching</td>
<td>0.059</td>
<td>0.072</td>
<td>0.076</td>
</tr>
<tr>
<td>Polynomial</td>
<td>MLE</td>
<td>0.059</td>
<td>0.072</td>
<td>0.062</td>
</tr>
<tr>
<td>All Lags</td>
<td>Moment-Matching</td>
<td>0.047</td>
<td>0.073</td>
<td>0.065</td>
</tr>
<tr>
<td>All Lags</td>
<td>MLE</td>
<td>0.052</td>
<td>0.078</td>
<td>0.065</td>
</tr>
</tbody>
</table>

similar results. However, the polynomial controls suggest larger teacher effects on test scores four years out, a result that holds for each estimator. Appendix Figure A.8 shows that this is also true of ELA teachers (naturally, because only fourth grade teachers have students with four-year-later test scores, and fourth grade teachers generally teach both ELA and math). The reason for this discrepancy is unclear.

1.7 Conclusion

The results of this paper have two policy implications that are in tension. First, we might want to incorporate teacher effects on attendance into teacher value-added measures. Teachers have substantial effects on attendance, which is valuable in itself. And since teacher effects on attendance are approximately as large and persistent as teacher effects on test scores and more predictive of graduation, we might favor attendance as a useful proxy for teacher effects on student well-being. On the other hand, teacher effects on the most plausibly welfare-relevant outcomes, long-term test scores and high school graduation, are poorly captured by teacher effects on the immediately observable outcomes, test scores and attendance. If improving short-term and long-term outcomes are very different tasks, then causing teachers to improve short-term outcomes might distract them from more important tasks.

First, the case for measuring teacher effects on attendance. Absenteeism is pervasive in urban school districts, and while teachers explain only a small portion of the variance
in student attendance, they can be effective in reducing absenteeism. A teacher one standard deviation above average at increasing attendance teaching a class of 25 students increases attendance by 78 student-days\(^5\). Holding teacher behavior constant, incorporating attendance into existing value-added measures would make these measures more fair and informative, since teachers who are effective in improving test scores are not especially likely to improve attendance. Since school districts routinely collect attendance data and since many districts have adopted quantitative value-added measures, it would be very feasible for these districts to incorporate attendance and other outcomes into value-added scoring.

However, there may be risks to compensating, evaluating, or firing teachers based on attendance, analogous to the risks in incentivizing high test scores. Although there is a wide literature on incentivizing schools for higher test scores or meeting proficiency standards, the effects of high-stakes incentives for individual teachers are far from clear. Several studies discuss group incentives: Fryer (2013) discusses a randomized trial in which New York City public schools were eligible for more funding if they reached test score targets, and schools usually chose group incentives. Test scores did not improve in treatment schools. On the other hand, Lavy (2002) studied group incentives on several performance measures, including test scores, in Israel and found that they increased both contemporaneous test scores and a variety of outcome measures in the following year. Although incentives for higher test scores are intended to increase teacher effort, teachers often seem to respond in less desirable ways, such as increasing time spent on test prep (Glewwe et al., 2010). And even in the presence of school-level performance incentives that only weakly affect individual teachers, teachers may “teach to the test” (Klein et al., 2000) or cheat (Jacob and Levitt, 2003, Jacob, 2005, Loughran and Comiskey, 1999).

Although I show that high attendance value-added teachers have historically improved their students’ achievement, this does not imply that teachers should be compensated,

\(^5\)This is true for the elementary and middle school students in the analysis sample, in which the variance of teacher effects on attendance is 0.069, and the standard deviation of attendance in the analysis sample is 11.95 days.
evaluated, or fired based on attendance-based value-added measures. Similarly, this paper has little to say about whether score-based value-added should be a component of teacher evaluation. Policymakers face a multitasking problem in the spirit of Holmstrom and Milgrom (1991): we want teachers to make their students motivated, persistent, and informed, but we can only design contracts on the basis of observable factors. The risk of perverse incentives that comes with test score-based teacher evaluation measures may make attendance-based value added more appealing as an alternative, but it should also caution us that incentivizing teachers for student behavior may lead to unintended outcomes. For example, teachers could encourage students to come to school even when sick, or make class more fun at the expense of being edifying by, for example, showing movies. In addition, in the long run, teachers may require significantly higher pay to compensate them for the stress and uncertainty that a merit-based pay and retention system could generate.

Many questions remain unanswered in this area. Using this data, it is possible to explore heterogeneity in teacher effects; for example, do teachers have larger effects on absences for students who are absent more often? It would also be helpful to track these students farther into the future to explore the effects of teachers who reduce absences on high school graduation rates, college attendance and completion rates, and income.
Chapter 2

Measuring the Distribution of (Teacher) Value-Added

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., 2014a; 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%. \footnote{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., 2014a).}

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 $\mu$ are drawn identically and independently from the same distribution, $\mu_j \sim F$, and the distribution $F$ itself is of interest. The individual components of $\mu$ may also be of interest. High-dimensional covariates and few observations for each teacher are common complications, making it inadvisable to simply estimate $\mu$ 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 $\mu$. 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 2.1 I develop a toy model to motivate why policymakers may care about the variance of teacher effects. In Section 2.2 I recap the historical development of the value-added literature and the settings in which value-added estimators have been used. Section 2.3 describes several estimators whose properties I develop and compare. In particular, subsection 2.3.4 notes that when some covariates do not vary within teacher, the variance of teacher effects is only partially identified without further assumptions, and subsection 2.3.5 discusses how different estimators behave in the presence of errors in variables. Section 2.4 discusses the behavior of several procedures in Monte Carlo experiments, and Section 2.5 concludes with recommendations about which estimator to use.
2.1 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 $\sigma_\mu$.

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

$$
E[\mu_j] - E[\mu_j | F(\mu_j) < p] = 0 - E[\mu_j | F_\mu(\mu_j) < p]
$$

$$
= -\sigma_\mu E[\varepsilon_j | F_\varepsilon(\varepsilon_j) < p]
$$

$$
\propto \sigma_\mu.
$$

2.2 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). 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 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, $\hat{\text{Var}}(\mu_j)$, 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 $\hat{\text{Var}}(\mu_j)$. The literature has typically interpreted forecast bias as a sign of insufficient controls for student-teacher sorting, but it can also reflect bias in $\hat{\text{Var}}(\mu_j)$, an issue I consider in this paper. Randomized and quasi-experimental analyses have somewhat ameliorated concerns that sorting on unobservables biases estimates

---

\footnote{Briggs and Domingue 2011 finds that teachers’ educational backgrounds do predict teacher effects}
Table 2.1: Estimates of the variance of teacher effects, $\text{Var}(\mu_j)$, 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.

<table>
<thead>
<tr>
<th></th>
<th>Math</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Var}(\mu_j)^{1/2}$</td>
<td>0.219</td>
<td>0.175</td>
</tr>
<tr>
<td>1 - forecast bias</td>
<td>0.905</td>
<td>1.089</td>
</tr>
<tr>
<td></td>
<td>[0.552, 1.258]</td>
<td>[0.523, 1.655]</td>
</tr>
</tbody>
</table>

of the variance of teacher quality upwards. 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. 2014a).

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 2.1, show that a teacher one standard deviation above average improves 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.

3 In addition to the studies by Kane and Staiger (2008) and Chetty et al. (2014a) 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.
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. (2014a) 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.

### 2.3 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. (2014a), "modified-KS." Both maximum (quasi-)likelihood and modified-KS consistently estimate this model. I show that the Kane and Staiger estimator consistently estimates this model after imposing a no-sorting restriction, and is otherwise negatively biased.

Observations are at the student level. Student $i$ has classroom $c(i)$, teacher $j(i)$, test score $y_{it}$, and covariates $x_{it}$. Data is drawn from some distribution $D$. (Although a likelihood function will be derived using normality assumptions, $D$ 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

---

<table>
<thead>
<tr>
<th>Consistent under baseline model</th>
<th>MLE</th>
<th>Bias-Corrected MLE</th>
<th>Kane and Staiger</th>
<th>Mod-KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent under baseline + no sorting</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Sign of bias under baseline model</td>
<td>Up</td>
<td>?</td>
<td>Down</td>
<td>?</td>
</tr>
<tr>
<td>Closed-form solution</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Closed-form asymptotic standard errors</td>
<td>MLE</td>
<td>?</td>
<td>GMM</td>
<td>GMM</td>
</tr>
</tbody>
</table>

---

4I use bolded letters (i.e. $x$) to represent vectors, and bolded and italicized letters (i.e. $x$) to represent matrices.
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 \( y_j \in \mathbb{R}^{n_s} \), a matrix \( x_j \in \mathbb{R}^{n_s \times k} \), and mean covariates \( \bar{x}_j \in \mathbb{R}^k \). \( y_j \) and \( x_j \) 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

\[
y_j \equiv \mu_j + x_j \beta + \nu_j. \tag{2.1}
\]

The teacher effect, \( \mu_j \), 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 \( \nu_j \) is orthogonal to covariates \( x_j \), so that we consistently estimate \( \beta \). 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 \( \nu_j \perp s_j|x_j, \bar{x}_j \). To see why this second restriction is necessary, imagine that all teachers are identical – \( m_j = 0 \) – but some teachers are consistently assigned students with high values of \( \nu_j \). In that case, some teachers will consistently appear to have students that over- or under-perform what would be expected from their covariates, making it appear that teachers vary in their quality when they actually do not.

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

\[
\mu_j = \bar{x}_j^T \lambda + \bar{\mu}_j, \quad \bar{\mu}_j \perp \bar{x}_j, s_j \tag{2.2}
\]

\( \lambda \) is a vector governing the association of covariates with teacher quality. It could capture

\[\bar{x}_j \text{ is a precision-weighted mean, in a way that will be made clear.}\]
teacher-specific characteristics – for example, more experienced teachers are better – or reflect sorting – for example, teachers of honors classes may be better.

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

Combining Equations 2.1 and 2.2, $E_D[y_j|I_n \otimes \text{vec}(x_j), \tilde{x}_j] = x_j \beta + \tilde{x}_j^T \lambda$. When there are covariates that do not have within-teacher variation, $$\beta$$ and $$\lambda$$ are only partially identified. Section 2.3.4 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 2.3.1 through 2.3.3, I assume that all covariates have within-teacher variation so that the model is point identified.

### 2.3.1 Maximum (Quasi-)Likelihood

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

$$E_D^* [y_j|I_n \otimes \text{vec}(x_j), \tilde{x}_j] = (I_n \otimes \text{vec}(x_j)) \pi + \tilde{x}_j \lambda.$$  

We need that $$(I_n \otimes \text{vec}(x_j)) \pi = x_j \beta$$. Finally, let’s put more structure on the covariance of errors and assume homoskedasticity with respect to $$s_j$$. Define $E [\nu_j \nu_j^T | s_j] = \Sigma_j$. Denote parameters $$\eta = (\beta, \lambda, \sigma^2_\mu, \sigma^2_\theta, \sigma^2_\varepsilon)$$.

$$\Sigma (\eta, x_j, s_j)_{i,i'} = \sigma^2_\mu + \sigma^2_\theta + \sigma^2_\varepsilon \text{ when } i = i'$$

$$\Sigma (\eta, x_j, s_j)_{i,i'} = \sigma^2_\theta + \sigma^2_\varepsilon \text{ when } i \neq i' \text{ but } i \text{ and } i' \text{ are in the same class}$$

$$\Sigma (\eta, x_j, s_j)_{i,i'} = \sigma^2_\mu \text{ when } i \text{ and } i' \text{ are not in the same class}$$

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 \( y_j \) and \( m_j \). Appendix C.1 proves the validity of a quasi-likelihood interpretation: maximum likelihood based on normality delivers consistent estimates of \( \eta \), even when the true distribution \( D \) does not have normal disturbances. If the true values of parameters are \( \eta^* \) and maximum likelihood based on normality assumptions estimates \( \eta_F \), then \( \eta^* = \eta_F \). That is, consider the model

\[
y_j | x_j, \bar{x}_j, s_j \sim N \left( x_j \beta + \bar{x}_j^T \lambda, \sum (\eta^*, s_j) \right)
\]

with the corresponding likelihood function \( f \left( y_j, x_j, \bar{x}_j, s_j ; \eta \right) \). Appendix C.1 proves that \( \eta_F = \arg \max_{\eta} E_D \log f \left( y_j, x_j, \bar{x}_j, s_j ; \eta \right) \). Appendix C.2 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 \( h_c = \frac{1}{\sigma^2_c + \sigma^2_t / |I(c)|} \), precision-weighted teacher-level means \( \bar{y}_j \), and classroom-level deviations from teacher means \( \tilde{y}_c \). Equations C.10 through C.14 define these terms.

Solving for the likelihood without integrating out teacher effects, as in Appendix Equation C.18 gives an integral with a Bayesian interpretation that yields an Empirical Bayes posterior: Teacher effects are drawn \( \mu_j \sim N \left( \bar{x}_j^T \lambda, \sigma^2_{\mu} \right) \), and test scores are drawn \( \tilde{y}_j \sim N \left( \mu_j + \bar{x}_j^T \beta, \frac{1}{\sigma^2_c + \sum_{c \in C(j)} h_c} \right) \), so the Empirical Bayes posterior of teacher \( j \)'s value-added is

\[
\mu_j \sim N \left( \frac{\sigma^2_{\mu}}{\sigma^2_{\mu} + 1 / \sum_{c \in C(j)} h_c} (\tilde{y}_j - \bar{x}_j^T \beta) + \frac{1}{\sigma^2_{\mu} + 1 / \sum_{c \in C(j)} h_c} \bar{x}_j^T \lambda, \sigma^2_{\mu} + \sum_{c \in C(j)} h_c \right) \left( \frac{1}{\sigma^2_{\mu}} + \sum_{c \in C(j)} h_c \right)^{-1}
\]

(2.3)

The solutions for \( \hat{\beta} \) and \( \hat{\lambda} \) are intuitive. After concentrating out \( \hat{\lambda} \), \( \hat{\beta} \) attempts to jointly minimize students’ deviations from the classroom mean and classrooms’ deviations from the teacher mean:
\[ \hat{\beta} = \arg \min_{\beta} \frac{1}{\sigma^2} \sum_j \sum_{c \in C(j)} \sum_{i \in I(c)} \left( \bar{y}_i - \bar{x}_i^T \beta \right)^2 + \sum_j \sum_{c \in C(j)} \tilde{h}_c \left( \bar{y}_c - \bar{x}_c^T \beta \right)^2 \] (2.4)

\[ \hat{\lambda} = \arg \min_{\lambda} \sum_j \left( \frac{1}{\sum_c \tilde{h}_c} + \sigma^2 \right)^{-1} \left( \bar{y}_j - \bar{x}_j^T \hat{\beta} - \bar{x}_j^T \lambda \right)^2 \] (2.5)

When there are no classroom-level shocks, \( \hat{\beta} \) and \( \hat{\lambda} \) coincide with the estimands from a correlated random effects framework. When \( \sigma^2 = 0 \), precisions \( h_c \) are proportional to the number of students in the class, so each observation is given equal weight. Equation (2.4) collapses to

\[ \hat{\beta} = \arg \min_{\beta} \sum_i \left( y_i - \bar{y}_{j(i)} - \left( x_i - \bar{x}_{j(i)} \right)^T \beta \right)^2, \]

and Equation (2.5) becomes

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

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

\[ y_i = x_i^T \beta + \bar{x}_{j(i)}^T \lambda + \epsilon_i. \]

**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 C.5 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.
Bias correction to variance of teacher effects

The quantity of interest is
\[ \text{Var}(\mu_j) = \text{Var}(\hat{x}_j^T \lambda) + \sigma^2_\mu. \]

An obvious estimator is the sample analog:
\[ \hat{\text{Var}}(\mu_j|\lambda) = \frac{1}{n} \sum_j (\hat{x}_j^T \lambda)^2 - \left( \frac{1}{n} \sum_j \hat{x}_j^T \lambda \right)^2 + \hat{\sigma}^2_\mu. \]

However, the sample analog is biased upwards. \( \mathbb{E} \left[ \text{Var}(\hat{x}_j^T \lambda|\lambda) \right] > \mathbb{E} \left[ \text{Var}(\hat{x}_j^T \lambda) \right] \), for a clear reason: estimation error in \( \hat{\lambda} \) will tend to make this quantity larger. Imagine that \( \lambda = 0 \): \( \hat{\lambda} \) 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 C.19 the sample analog is biased upwards by
\[ \mathbb{E} \left[ \text{Var}(\hat{x}_j^T \lambda|\lambda) \right] - \text{Var}(\hat{x}_j^T \lambda) = \mathbb{E} \left[ (\hat{x}_j - \mathbb{E} \hat{x}_j)^T \text{Cov}(\hat{\lambda}) (\hat{x}_j - \mathbb{E} \hat{x}_j) \right]. \] (2.6)

Therefore, a bias-corrected estimator of the variance of teacher effects is
\[ \hat{\text{Var}}(\mu_j) = \frac{1}{n} \sum_j (\hat{x}_j^T \lambda)^2 - \left( \frac{1}{n} \sum_j \hat{x}_j^T \lambda \right)^2 + \hat{\sigma}^2_\mu - \frac{1}{n} \sum_j \left( \hat{x}_j^T - \frac{1}{J} \sum_k \hat{x}_k \right)^T \hat{\Sigma}_\lambda \left( \hat{x}_j^T - \frac{1}{J} \sum_k \hat{x}_k \right). \] (2.7)

where \( \hat{\Sigma}_\lambda \) is the asymptotic variance of \( \hat{\lambda} \).

2.3.2 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. (2014a). As discussed in Section 2.2, 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 2.3.2 shows that although the estimator is consistent when teacher effects are not correlated with covariates – \( \lambda = 0 \) – this estimator is asymptotically downward biased when \( \lambda \neq 0 \). Section 2.3.3 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 2.3.1 under only a sorting on observables requirement.

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: \( \beta, \sigma^2_\mu, \sigma^2_\beta, \) and \( \sigma^2_\varepsilon \). Then we estimate each teacher’s value of \( \mu_j \) 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 \( \hat{\beta} \), then we use \( \hat{\beta} \) to generate residuals. Next, we use a “moment-matching” procedure to estimate the variances \( \sigma^2_\mu, \sigma^2_\beta, \) and \( \sigma^2_\varepsilon \) based on variances and covariances of residuals. In more detail:

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

\[
\hat{\beta} = \arg \min_b \sum_i \left( y_i - x_i^T b \right)^2 \\
= \beta + \left( \sum x_i x_i^T \right)^{-1} \sum x_i \left( \mu_j(i) + \nu_i \right) \\
\mathbb{E} \left[ \hat{\beta} \right] = \beta + \left( \mathbb{E} \left[ x_i x_i^T \right] \right)^{-1} \mathbb{E} \left[ x_i \mu_j(i) \right]
\]

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

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

$$\hat{\sigma}^2_m = \frac{2}{\sum_j |C(j)| (|C(j)| - 1)} \sum_j \sum_{c,c' \in C(j), c \neq c'} \left( \hat{y}_c - \bar{x}_c^T \hat{\beta} \right) \left( \hat{y}_{c'} - \bar{x}_{c'}^T \hat{\beta} \right)$$ (2.8)

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

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

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

$$\hat{\sigma}^2_\theta = \text{Var}(y_i - x_i^T \hat{\beta}) - \hat{\sigma}^2_\mu - \hat{\sigma}^2_\epsilon.$$

**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)](https://www.jstor.org/stable/2685028), or through the Generalized Method of Moments.

Denote the parameters $\eta = (\beta, \sigma^2_\mu, \sigma^2_\epsilon, \sigma^2_\theta)$. The moment function, which is at the teacher level, is

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

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

\[
\hat{\beta}^n = \left( \sum_i \omega^n_{j(i)} x_i x_i^T \right)^{-1} \sum_i \omega^n_{j(i)} x_i y_i
\]

\[
\hat{\sigma}_{\mu}^2(n) = \frac{1}{\sum_j \omega^n_j |C(j)|} \sum_{j} \sum_{c \in C(j), c' \in C(j), c' \neq c} \omega^n_j (\hat{y}_c - \bar{x}_c^T \hat{\beta}^n) (\hat{y}_c - \bar{x}_c^T \beta^n)
\]

\[
\hat{\sigma}_{\epsilon}^2(n) = \frac{1}{\sum_j \omega^n_j (|I(c)| - 1)} \sum_i \omega^n_{j(i)} (\hat{y}_i - \hat{x}_i^T \hat{\beta}^n)^2
\]

\[
\hat{\sigma}_{\theta}^2(n) = \frac{1}{\sum_j \omega^n_j (|I(c)|)} \sum_i \omega^n_{j(i)} (y_i - \hat{x}_i^T \hat{\beta}^n)^2 - \hat{\sigma}_{\mu}^2(n) - \hat{\sigma}_{\epsilon}^2(n)
\]

**Individual Teacher Effects**

Although a teacher’s mean residuals are an unbiased estimate of \( \mu_j \), Kane and Staiger use shrinkage to produce a best linear predictor of \( \mu_j \). First, generate the precision \( h_c = \text{Var}(\hat{y}_c - \bar{x}_c^T \beta) \) of each mean classroom residual; these are the same precisions used for maximum likelihood in Equation C.13. Then construct a precision-weighted mean using \( h_c \) and multiply it by shrinkage factor \( \rho_j \) to generate a mean squared error-minimizing estimate of \( \mu_j \):

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

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

(2.9)

Kane and Staiger note that when \( \mu, \theta, \) and \( \epsilon \) are normally distributed, Equation 2.9 has a Bayesian interpretation. The estimated variances \( \hat{\sigma}_{\mu}^2, \hat{\sigma}_{\theta}^2, \) and \( \hat{\sigma}_{\epsilon}^2 \) 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 2.9 equals Equation 2.3 from maximum likelihood, when \( \lambda = 0 \): Conditional on parameter estimates, both procedures deliver the same estimated individual teacher effects. However, even with the imposition of \( \lambda = 0 \), 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.

Inconsistency and bias under misspecification

Consistency and bias of \( \hat{\sigma}^2_p \)

As discussed extensively in Guarino et al. (2014) and mentioned in Chetty et al. (2014a), 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 \( \hat{\beta} \) will be biased when estimated in a regression that omits teacher fixed effects; here I demonstrate that omitted variable bias in \( \hat{\beta} \) leads to an asymptotic negative bias in \( \hat{\text{Var}}(\mu_j) \). Intuitively, variation in teacher effects that is correlated with student characteristics is incorrectly attributed to the student characteristics. Equation 2.8 gives

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

Appendix C.5 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:

\[
-3 \leq \frac{\text{Bias}(\hat{\text{Var}}(\mu_j))}{\frac{\sum_j \bar{x}_j^T \mu_j}{N_{\text{teachers}}} \left( \frac{\sum_j x_j x_j^T}{N_{\text{students}}} \right)^{-1} \frac{\sum_j \bar{x}_j \mu_j}{N_{\text{teachers}}}} \leq -1 \\
\text{(2.10)}
\]

Equation 2.10 makes several facts apparent. The estimator is always negatively biased even asymptotically, and bias is more severe when sorting is strong. This happens when \( \lambda \) is large in magnitude.
2.3.3 Modified version of above estimator: Modified-KS

As discussed above, omitted variables bias parameter estimates in the Kane and Staiger estimator. Chetty et al. (2014a) suggest remedying this by including teacher fixed effects when residualizing. That is, we obtain \( \hat{\beta} \) as the coefficient on \( x_i \) in a regression of outcomes on \( x_i \) and teacher fixed effects. 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 (\( \theta_c \)). 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 \( \hat{\beta} \rightarrow_p \beta \), which in turn implies that \( \text{Var}(\mu_j) \rightarrow_p \text{Var}(\mu_j) \). However, Section 2.3.5 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.

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 \( \hat{\beta} \) is now estimated off of within-teacher variation:

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

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

6 Chetty et al. (2014a) 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 \( \hat{\beta} \) but do not study the rest of their model.
2.3.4 Covariates without within-teacher variation

Analysis up to this point has assumed that it is possible to separately identify $\beta$ and $\lambda$. This model is not point identified when a column of $x_j$ is constant within each teacher. In that case, adding a constant to the corresponding component of $\beta$ and subtracting it from the corresponding component of $\lambda$ would yield the same prediction of $y_j$. 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 $\psi_i$.  

To see the effects of introducing covariates that do not vary within teacher, recall that

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

---

7Randomly assigning students to teachers would identify that either $\beta^k$ or $\lambda^k$ 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 $\lambda$ corresponding to students’ previous test scores would be zero: there would be no association between teacher quality and student attributes. And the component of $\beta$ corresponding to teacher gender would be zero: There would be no relationship between teacher gender and student quality.
the original model is given by

\[ y_j = \mu_j + x_j \beta + v_j, \quad v_j \perp 1, x_j \]

\[ \mu_j = \tilde{x}_j^T \lambda + \tilde{\mu}_j, \quad \tilde{\mu}_j \perp 1, \tilde{x}_j \]

\[ v_j \perp s_j | x_j, \tilde{x}_j \text{ sorting on observables} \]

The predicted value of \( y_j \) is

\[ E^* [y_j | x_j, \tilde{x}_j, s_j] = x_j \beta + \tilde{x}_j^T \lambda \]

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 \( x_j \) as \( x_j = (x^0_j, x^1_j \ell, \ldots, x^{K_1}_j \ell) \), where each column of \( x^0_j \in \mathcal{R}^{n_s \times (K-K_1)} \) has within-teacher variation. Then the following equations are equally consistent with the data for all \( \alpha \in \mathcal{R}^{K_1} \):

\[ E^*_D \left[ y_j | \mu_j, x^0_j, x^1_j, \ldots, x^{K_1}_j, s_j \right] = \mu_j + x^0_j \tilde{\beta} + \sum_{k=1}^{K_1} \alpha_k x^k_j \gamma_k \]

\[ \equiv \mu_j + x_j \beta, \quad \beta^T = (\tilde{\beta}^T, \alpha_1 \gamma_1, \ldots, \alpha_{K_1} \gamma_{K_1}) \]

\[ E^*_D \left[ \mu_j | \tilde{x}^0_j, x^1_j, \ldots, x^{K_1}_j, s_j \right] = \tilde{x}^0_j \tilde{\lambda} + \sum_{k=1}^{K_1} (1 - \alpha_k) x^k_j \gamma_k \]

\[ \equiv \tilde{x}_j^T \lambda, \quad \lambda^T = (\tilde{\lambda}^T, (1 - \alpha_1) \gamma_1, \ldots, (1 - \alpha_{K_1}) \gamma_{K_1}) \quad (2.11) \]

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

\[ \text{Var} (\mu_j) = \text{Var} (\tilde{x}_j^T \lambda) + \sigma^2_{\mu} \]

\[ = \text{Var} \left( \tilde{x}_j^T \lambda + \sum_{k=1}^{K_1} (1 - \alpha_k) x^k_j \gamma_k \right) + \sigma^2_{\mu} \]

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

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

**Maximum (Quasi-)Likelihood**

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

First, set the last \( K_1 \) components of \( \hat{\beta} \) to zero, setting \( \alpha = (0, \ldots, 0) \), 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 \( \alpha \): Set \( \hat{\beta} \) to the first \( K - K_1 \) components of \( \hat{\beta} \), \( \hat{\lambda} \) 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 \( \tilde{\text{Var}}(\mu_j)(\alpha) \) for any \( \alpha \) as

\[
\tilde{\text{Var}}(\mu_j) = \sigma^2_{\mu} + \tilde{\text{Var}} \left( x_j^0 \tilde{\lambda} + \sum_{k=1}^{K_1} (1 - \alpha_k)x_j^k \gamma_k \right).
\]

By varying \( \alpha \), we can see that there are various values of \( \text{Var}(\mu_j) \) that are consistent with the point-identified parameters \( \tilde{\lambda} \) and \( \gamma \).
First, in the case that $\alpha$ consists of all zeros, any variation that could potentially be attributed to either teachers ($\lambda$) or students ($\beta$) is attributed to teachers. In the case that $\alpha$ 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 $\alpha$ so that $\sum_{k=1}^{K_1} (1 - a_k)^2 x_j^k \gamma_k$ cancels out as much of the variation in $x_j^\alpha \lambda$ as possible. And as $\alpha$ 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 $v_i$. Restricting each component of $\alpha$ 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.

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 $\alpha$ is zero. Then $\hat{\beta}$ equals the first $K - K_1$ components of $\hat{\beta}$. However, since the modified-KS estimator does not explicitly model sorting, $\gamma$ is unknown. Therefore, we cannot find $\text{Var}(\mu_j)$ as a function of $\alpha$, but rather as a function of $\phi \equiv (\alpha_1 \gamma_1, \ldots, \alpha_{K_1} \gamma_{K_1})^T$, which is less easy to interpret.
\[
\text{Var}(\mu_j)_{\phi=\text{mod-KS}} = \frac{1}{\sum_j |C(j)| (|C(j)| - 1)} \sum_j \sum_{c \in C(j), c' \notin C(j)} \left( g_c - \hat{x}_c^T \hat{\beta} - x_j^T \phi \right) \left( g_{c'} - \hat{x}_{c'}^T \hat{\beta} - x_j^T \phi \right)
\]

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

\[
\text{arg min}_{\phi} \text{Var}(\mu_j)_{\phi=\text{mod-KS}} = \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right)^{-1} \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right) x_j^0 \left( g_c - \hat{x}_c^T \hat{\beta} \right) \left( g_{c'} - \hat{x}_{c'}^T \hat{\beta} \right)
\]

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

\[
\text{plim} \left( \text{arg min}_{\phi} \text{Var}(\mu_j)_{\phi=\text{mod-KS}} \right) = \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right)^{-1} \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right) x_j^0 \left( g_c - \hat{x}_c^T \hat{\beta} \right) \left( g_{c'} - \hat{x}_{c'}^T \hat{\beta} \right)
\]

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

\[
\text{arg min}_{\phi} \text{Var}(\mu_j)_{\phi=\text{mod-KS}} = \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right)^{-1} \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right) x_j^0 \left( g_c - \hat{x}_c^T \hat{\beta} \right) \left( g_{c'} - \hat{x}_{c'}^T \hat{\beta} \right)
\]

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

\[
\text{plim} \left( \text{arg min}_{\phi} \text{Var}(\mu_j)_{\phi=\text{mod-KS}} \right) = \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right)^{-1} \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right) x_j^0 \left( g_c - \hat{x}_c^T \hat{\beta} \right) \left( g_{c'} - \hat{x}_{c'}^T \hat{\beta} \right)
\]

\[
\text{plim} \left( \text{arg min}_{\phi} \text{Var}(\mu_j)_{\phi=\text{mod-KS}} \right) = \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right)^{-1} \left( \sum_j \sum_{c \in C(j), c' \notin C(j)} \sum c' \neq c \right) x_j^0 \left( g_c - \hat{x}_c^T \hat{\beta} \right) \left( g_{c'} - \hat{x}_{c'}^T \hat{\beta} \right)
\]
2.3.5 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. \( \mathbf{y}_j \in \mathbb{R}^{n_i} \), \( \mathbf{z}_j \in \mathbb{R}^{n_i} \), \( \mathbf{x}_j \in \mathbb{R}^{n_i} \), and \( \beta, \lambda, \bar{x}_j \), and \( \bar{z}_j \) are scalars. The data is generated by

\[
\mathbf{y}_j = \mathbf{u}_j + \mathbf{z}_j \mathbf{\beta} + \mathbf{\nu}_j \\
\mathbf{u}_j = \bar{\mathbf{z}}_j \mathbf{\lambda} + \bar{\mathbf{\mu}}_j
\]

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

The Kane and Staiger estimator estimates

\[
\hat{\beta}^{KS} \to_p \frac{\text{Cov}(\mathbf{y}_i, \mathbf{x}_i)}{\text{Var}(\mathbf{x}_i)}
\]

\[
= \frac{\text{Cov}(\bar{\mathbf{z}}_{(i)}, \mathbf{z}_i) \lambda + \text{Var}(\mathbf{z}_i) \mathbf{\beta}}{\text{Var}(\mathbf{x}_i)}
\]

\[
= \frac{\text{Var}(\mathbf{z}_i)}{\text{Var}(\mathbf{x}_i)} \beta + \frac{\text{Var}(\bar{\mathbf{z}}_j)}{\text{Var}(\bar{\mathbf{x}}_j)} \lambda
\]

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, \( \text{Var}(\mathbf{u}_j) \) was biased downwards; now its bias cannot be signed.

The modified-KS estimator uses within-teacher variation:

\[
\hat{\beta} \to_p \frac{\text{Var}(\mathbf{z}_i \text{ demeaned})}{\text{Var}(\mathbf{x}_i \text{ demeaned})} \beta
\]

68
Calibration

We can calibrate $\sigma$ 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 $\sigma^2 = 0.2$. 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,

$$z_i = \bar{z}_{j(i)} + \bar{z}_c(i) + \bar{z}_i$$

$$\text{Var}(\bar{z}_j) = 0.34$$

$$\text{Var}(\bar{z}_c) = 0.04$$

$$\text{Var}(\bar{z}_i) = 0.42$$

$$x_i = z_i + \sigma \varepsilon_i$$

$$\sigma^2 = 0.2$$

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

$$\lim_{N_{\text{teachers}} \to \infty, N_{\text{students per teacher}} \to \infty} \hat{\beta}^{\text{KS}} = 0.8 \beta$$

$$\lim_{N_{\text{teachers}} \to \infty, N_{\text{students per teacher}} \to \infty} \hat{\beta}^{\text{mod-KS}} = \frac{\text{Var}(z_i) - \text{Var}(\bar{z}_j)}{\text{Var}(x_i) - \text{Var}(\bar{z}_j)} \beta$$

$$= 0.70 \beta$$

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
2.4 Monte Carlo Experiments

2.4.1 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 2.3.

| Table 2.3: Summary statistics, New York City public schools. |
|-----------------|----------|---------|------|-----|---------|
| 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
Figure 2.1: $\text{Var}(\mu_j)$: Estimates from partitioning the data by number of years taught, in solid lines, and from whole data, in dotted horizontal lines.
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 2.1 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.

### Table 2.4: Summary statistics, New York City public schools: students of math teachers who appear in the data for three years.

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

In order to ensure that the estimates are comparable, all further analysis proceeds on 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 2.4 gives summary statistics for this reduced dataset. Table 2.6 shows estimates of $\hat{\text{Var}}(\mu_j)$ for each estimator for the effects of math teachers on math scores, among teachers who appear in the data in three years. The
Table 2.5: Estimates of $\hat{\text{Var}}(\mu_j)$ for the effects of math teachers on math scores, for teachers who can be linked to students.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\sigma}^2_\mu$</th>
<th>$\hat{\text{Var}}(\hat{\lambda}_j)$</th>
<th>$\hat{\text{Var}}(\mu_j)$</th>
<th>CI</th>
<th>CI (Robust)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kane and Staiger</td>
<td>3.46%</td>
<td></td>
<td></td>
<td>[3.46%, 3.47%]</td>
<td></td>
</tr>
<tr>
<td>Mod-KS</td>
<td>4.52%</td>
<td></td>
<td></td>
<td>[4.52%, 4.53%]</td>
<td></td>
</tr>
<tr>
<td>MLE</td>
<td>3.15%</td>
<td>2.24%</td>
<td>5.39%</td>
<td>[5.39%, 5.40%]</td>
<td>[5.31%, 5.47%]</td>
</tr>
<tr>
<td>Bias-Corrected MLE</td>
<td>3.15%</td>
<td>2.22%</td>
<td>5.37%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias-Corrected MLE (Robust)</td>
<td>3.15%</td>
<td>0.00%</td>
<td>3.15%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 2.6: Estimates of $\hat{\text{Var}}(\mu_j)$ 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.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\sigma}^2_\mu$</th>
<th>$\hat{\text{Var}}(\hat{\lambda}_j)$</th>
<th>$\hat{\text{Var}}(\mu_j)$</th>
<th>CI</th>
<th>CI (Robust)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kane and Staiger</td>
<td>3.75%</td>
<td></td>
<td></td>
<td>[3.75%, 3.76%]</td>
<td></td>
</tr>
<tr>
<td>Mod-KS</td>
<td>5.24%</td>
<td></td>
<td></td>
<td>[5.24%, 5.25%]</td>
<td></td>
</tr>
<tr>
<td>MLE</td>
<td>3.16%</td>
<td>2.25%</td>
<td>5.42%</td>
<td>[5.41%, 5.42%]</td>
<td>[5.34%, 5.50%]</td>
</tr>
<tr>
<td>Bias-Corrected MLE</td>
<td>3.16%</td>
<td>2.20%</td>
<td>5.36%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias-Corrected MLE (Robust)</td>
<td>3.16%</td>
<td>0.02%</td>
<td>3.19%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
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 $\hat{\lambda}$.

2.4.2 Subsampling Experiments

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

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

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

Table 2.7 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 pro-
duces 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 $\hat{\text{Var}}(\mu_j)$ to the “true” answer $\text{Var}(\mu_j)$ allows for empirical estimates of the bias and variance of each estimator, as well as helping visualize the distribution. Figure 2.2 histograms draws of $\hat{\text{Var}}(\mu_j)$ from each estimator.

<table>
<thead>
<tr>
<th>Number of teachers</th>
<th>200</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>KS</td>
<td>92.0%</td>
<td>96.2%</td>
</tr>
<tr>
<td>Mod-KS</td>
<td>88.6%</td>
<td>94.7%</td>
</tr>
<tr>
<td>MLE</td>
<td>58.5%</td>
<td>71.2%</td>
</tr>
<tr>
<td>MLE (Robust)</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Table 2.7: Empirical coverage probabilities of confidence intervals based on asymptotic approximations over 1000 draws.

Figure 2.2: Histograms of estimates of $\hat{\text{Var}}(\mu_j)$ from subsamples.

When discussing how estimates from small subsamples relate to the “true” answer from
Figure 2.3: 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.

In 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 Figure 2.3 shows the “bias” of subsample estimates of $\text{Var}(\mu_j)$, 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.

2.5 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 \( (\sigma^2_p, \sigma^2_\beta, \sigma^2_\gamma) \) and coefficients \( (\beta, \lambda) \). 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 $\lambda = 0$ – 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, $\text{Var}(\mu_j)$ 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 $\alpha$ because it fully models the sorting process. Quasi-ML can generate an estimate of $\text{Var}(\mu_j)$ for $\alpha = (0, \ldots, 0)^T$, $\alpha = (1, \ldots, 1)^T$, and the value of $\alpha$ that minimizes $\text{Var}(\mu_j)$, while mod-KS can only give estimates for $\alpha = (0, \ldots, 0)$ or the value that minimizes $\text{Var}(\mu_j)$. Furthermore, in asymptopia quasi-ML gives a higher lower bound than mod-KS.
Chapter 3

Bureaucrat Effects on Local Economic Outcomes

Institutions are important. Development economists now widely agree that good governance and well-functioning institutions are among the central determinants of economic development (Acemoglu et al. (2005); Pande and Udry (2005); Baland et al. (2010)). Recent research has made considerable progress in unpacking the black box of governance to identify how specific institutional features – like political reservations for disadvantaged groups – impact government performance (Chattopadhyay and Duflo (2004); Bhavnani (2009); Besley and Burgess (2002)).

How much agency individuals retain in the presence of institutional constraints is an active area of research. Politicians appear to have significant discretion: lots of evidence, especially from India, shows that the identity of politicians – gender, religion, criminal status – matters for public good provision and economic activity (Jones and Olken (2005); Prakash et al. (2016); Clots-Figuereas (2012); Bhalotra et al. (2014); Iyer et al. (2012)). Much less is known about bureaucrats’ ability to exercise discretion. Most existing research has focused on bureaucrats’ relationship with politicians and the associated agency problems (Alesina and Tabellini (2007); Iyer and Mani (2012); Nath (2015)). There is a recent but fairly large

---

1With Jonas Hjort and Gautam Rao
literature on how institutional features and hiring practices affect the types of individuals the public sector attracts. But this selection literature typically takes the desirability of specific characteristics, such as honesty or academic ability, as given (Dal Bó et al. 2013; Hanna and Wang 2013; Ashraf et al. 2014); or, in the recent papers closest to ours, estimate how observable characteristics (e.g. age or high exam scores) affect bureaucrat performance (Bertrand et al. 2015; Bhavnani and Lee 2017)). Our contribution is to estimate the contribution of variation in bureaucrat efficacy to variation in local economic outcomes. Specifically, we model bureaucrat effects as the best linear predictor $\mu_j$ of a bureaucrat’s causal effect on some outcome. We use value-added methodology to estimate $\text{Var}(\mu_j)$ for District Collectors in India due to both observable and unobservable bureaucrat characteristics. We use several estimators from the teacher value-added literature. Although our estimates suggest that bureaucrat quality accounts for up to 7% of the variance in the value of large project completions and up to 2% of the variance in night light intensity, placebo and permutation tests demonstrate that those estimates are biased upwards and imprecise. We use randomization inference to demonstrate that we cannot reject the null hypothesis that District Collectors do not vary in their effects on project completions or night light intensity.

### 3.1 Background

The Indian Administrative Service is the “topmost” layer of the bureaucracy in India (Iyer and Mani 2012), responsible for running “all key government departments at the state and federal levels as well as a range of public sector enterprises and corporations (Bertrand et al. 2015). IAS officers are recruited either through India’s Civil Service Examination or, less often, through promotion from the lower state civil services. Entering the IAS through the Civil Services Examination is extremely difficult; in 2015, over 460,000 people began the Civil Service Exam and only 120 IAS slots were offered (Bertrand et al. 2015).

Direct recruits are assigned to state “cadres” by a complicated rule; allotment of recruits to states is governed by the state’s need for officers, a cap on the number serving in their
home state, reservation criteria for disadvantaged segments of society, and the number of top-ranked officers serving in a state \cite{Iyer and Mani 2012}. After five to ten years, an officer is promoted to District Officer (also known as District Collector, District Magistrate, or Deputy Commissioner). Our analysis focuses on District Officers, who are responsible for “ensuring law and order, providing certain judicial functions, organizing relief and rehabilitation in cases of natural disasters, implementing development policies, and overseeing all aspects of administration in a specific district.” \cite{Iyer and Mani 2012} There is one District Officer for each of India’s approximately six hundred districts.

District Officers are frequently transferred by state-level bureaucrats to different districts within the state, but transfers between states are very rare: Only 160 of 2,965 District Officers in our sample serve in more than one state. Our identification requires that these transfers do not lead to a correlation between bureaucrat quality and unobserved shocks to district-level outcomes. ²

### 3.2 Data

#### 3.2.1 Bureaucrat postings

Our main data source on IAS officer postings are records scraped from the Department of Personnel and Training (DoPT), under the central Ministry of Personnel, Public Grievances and Pensions. The DoPT provides, for each officer, a complete “executive record sheet”. The record sheet includes personal information, such as home state and languages spoken; education qualifications; state cadre assignment; training details and awards; and posting history, including the designation, rank, department, location, and tenure of each post an officer serves in.

²We require that there is no correlation between the “match effect” between a bureaucrat and a district and the bureaucrat’s quality, but this does not require that match effects are zero. For example, it is acceptable if bureaucrats are more often assigned to districts where they speak the local language and are more successful in those districts, as long as this is not more common for low- or high-quality bureaucrats. Similarly, it is acceptable if low-ability bureaucrats are transferred more often, as \cite{Iyer and Mani 2012} find, as long as they are not transferred to districts that are about to have negative or positive shocks, or to districts they are better matched to.
We restrict postings to only collector postings based on the designation of each post. Using the location of each officer’s posting, we are able to determine the bureaucrat’s assignment at the time. This data allows us to construct a monthly panel that identifies the bureaucrat’s district assignment in each month.

We discovered several data issues with the DoPT records, which we attempted to fix using Civil Lists, personnel records obtained from the Indian government. Although the Civil Lists are not available for the entirety of our sample, we used them as a supplement to fill in holes in the DoPT records. In some cases, a district had no collector listed in the DoPT records but did in the Civil Lists. Similarly, a bureaucrat sometimes had an employment gap listed in the DoPT records but none in the Civil Lists. In these cases, we assumed the Civil Lists were correct. In cases where a district appeared to have two officers in the DoPT records, we simply dropped the district for that time period.

Summary statistics are in Table 3.1. We find, consistent with Iyer and Mani (2012), that collectors are transferred quite frequently.

3.2.2 Project outcomes

We use measures of the status of projects that are compiled by the Centre for Monitoring the Indian Economy (CMIE), an economic think-tank based in Mumbai, in its proprietary CapEx dataset. The CapEx database consists of details of over 45,000 projects which create new productive capacity (whether in manufacturing or services) begun since 1996 across India. CMIE attempts to capture all projects with a capital cost of over 10 million rupees ($250,000), through a combination of monitoring the press, firm press releases, public tenders, trade publications, and even company websites. The projects include both private and public sector projects, and track project status and cost. The CapEx data indicates when each project

---

3Equivalent titles were deputy commissioner and district magistrate. We identified 10 different designations that were variants of these titles, and used these as our definition for a district collector. The designations we classified as Collectors are “Collector”, “Dy Commr”, “Collector & DM”, “Dist Magistrate”, “Dist Collector”, “D M & Collector”, “Dy Commr / D M”, “Dist Magistrate-cum-Dy Commr”, “DC & DM”, and “Dy Commissioner cum Dist Magistrate”. The following designations were not classified as collectors: “Dy Dev Commr”, “Dev Commr”, “Dy Collector”, “Dy Municipal Commr”, “Dy Dev Commr & C E O”, “Spl Collector”, “Addl Collector”, “Jt Collector”, “Sub Collector”.

82
Table 3.1: Summary statistics for the whole sample and for the largest connected set where each outcome is not null.

Notes. Full sample consists of all district collectors with a district assignment and serving after 1996. The “largest connected set” for an outcome refers to the subset of districts connected by district collector mobility where the outcome is not null, explained in detail in Section ??.

Number of district postings per collector is defined as the number of distinct districts a collector has been assigned to over his/her career. Tenure length is based off the begin and end dates of a posting in each officer’s record sheet (not the inferred begin date used when constructing the monthly panel).

was announced, was completed, became stalled, or was abandoned, a well as the capital expenditure on the project. These indicators are collected by CMIE as they follow-up every quarter with private-sector firms and public sector organizations to inquire about the status of their projects. We collapse this information at the district-month level so we can match project status with the identity of the serving district collector over the period 1996-2013. Most projects in our dataset involve a significant investments in physical infrastructure – building houses or generating electricity, for example.

Figures 3.1 plots the number of project completions per district-month, and 3.2 is a histogram of the number of project completions per posting. Note that both figures are on a log scale; the data is very skewed, with a large number of zeros.

### 3.2.3 Night lights

Night-time luminosity, which reflects electrification, is an increasingly common measure of economic activity. We use data from Defense Meteorological Satellite Program’s Operational
Linescan System on light intensity at night at the district level. It is not clear how to interpret the welfare impact of changes in electrification in the context of bureaucrat effects. If a district collector causes night electrification to increase by 1%, this could be because she increased economic activity by 1%, because she increased the availability of electricity directly, or even because she lobbied for electricity to be available for her district at the expense of another district.

Figure 3.3 plots night light intensity over time in India’s seven biggest states and in all of India. Generally, light intensity increases over time.
Figure 3.2: Histogram of the number of project completions per posting.

Figure 3.3: Night light intensity z-scores over time, in the seven biggest states and in all of India.
3.3 Bureaucrat Value-Added: Methodology

Value-added (VA) models have often been used by researchers to assess the variation in teachers’ effectiveness. For example, Chetty et al. (2014a) and Chetty et al. (2014b) found that teachers in a large American school district vary widely in their effect on test scores, and that teachers who increase test scores have large effects on their students’ outcomes later in life. These models are used in many states to assess the effectiveness of individual teachers.

We use several estimation strategies, outlined in subsection 3.3.3, to estimate the model described in subsection 3.3.1. 3.3.2 describes the variables we control for. In order to obtain exact p-values, we simulate the distribution of these estimators under the null hypothesis of no variation in bureaucrat quality by randomizing bureaucrat labels in a way similar to a permutation test. In addition to providing p-values, the simulations demonstrate that our estimators are biased upwards and have high dispersion, perhaps because our data is much smaller than is found in educational contexts. 3.3.4 describes our randomization inference strategy.

3.3.1 Model

Our model draws heavily on the model in Chapter 2 of this dissertation, which is in term similar to the model from Kane and Staiger (2008). Our bureaucrats are analogous to their teachers, our postings to their classrooms, and our observations to their students. Kane and Staiger’s model, in turn, derives from the literature on Empirical Bayes and on Hierarchical Linear Models. Bureaucrat effects are defined as the best linear predictor of their causal effect on some outcome of interest.

Denote the outcome as $y_{dt}$, where $d$ indexes districts and $t$ indexes time. The collector in district $d$ at time $t$ is $b(d, t)$. The outcome is the sum of the bureaucrat’s value-added $\mu_{b(d,t)}$ and covariates are $x_{dt}$. $s_b$ encapsulates the configuration of bureaucrats to districts. Also, stack all data from bureaucrat $b$ into outcomes $y_b$ and covariates $x_b$. The best linear
predictor of outcomes given bureaucrat value-added, covariates, and configuration is

\[ E^* [y_b | \mu_b, x_b, s_b] = \alpha + \mu_b + x_b \beta, \]  
(3.1)

and the best linear predictor of bureaucrat effects given the bureaucrat’s average covariates \( \bar{x}_b \) is

\[ E^* [\mu_b | \bar{x}_b, s_b] = \bar{x}_b^T \lambda. \]  
(3.2)

\( \lambda \) is a vector governing the association of covariates with bureaucrat quality. It could capture characteristics related to districts – for example, perhaps better bureaucrats are assigned to better districts – or it could capture characteristics related to bureaucrats: for example, perhaps female bureaucrats invest more in public goods, as in Clots-Figuereas (2012), Clots-Figueras (2011), and Chattopadhyay and Duflo (2004).

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

\[ E^* [y_b | I_n \otimes \text{vec}(x_b), \bar{x}_b] = (I_n \otimes \text{vec}(x_b)) \pi + \bar{x}_b \lambda. \]

We need that \( (I_n \otimes \text{vec}(x_b)) \pi = x_b \beta \), which rules out serial correlation.

Finally, we need to put more structure on the covariance of errors and assume homoskedasticity. Define \( \Sigma_{sb} \equiv \Sigma_{sb} \) and denote parameters \( \eta = (\alpha, \beta, \lambda, \sigma_{\mu}^2, \sigma_{\beta}^2, \sigma_{e}^2) \).

\[
\begin{align*}
\Sigma(\eta, x_b, s_b)_{(d,t),(d',t')} &= \sigma_{\mu}^2 + \sigma_{\beta}^2 + \sigma_{e}^2 & \text{when } & d = d', t = t' \\
\Sigma(\eta, x_b, s_b)_{(d,t),(d',t')} &= \sigma_{\beta}^2 & \text{when } & d = d', \ t \neq t' \\
\Sigma(\eta, x_b, s_b)_{(d,t),(d',t')} &= \sigma_{\mu}^2 & \text{when } & d \neq d'
\end{align*}
\]

\( \sigma_{\mu}^2 \) can be interpreted as the variance of the unpredictable component of bureaucrat effects, \( \sigma_{\beta}^2 \) as a posting-level shock, and \( \sigma_{e}^2 \) as a district-month specific shock.
3.3.2 Orthogonality Restrictions, Identification, and Controls

In order to ascribe a causal interpretation to parameter estimates, we need sorting on observables. To define this precisely, define errors for Equations 3.1 and 3.2:

\[
y_b = x_b \beta + x_b^T \lambda + v_b, \quad v_b \perp x_b, \, \bar{x}_b
\]
\[
\mu_b = \bar{x}_b^T \lambda + \bar{\mu}_b, \quad \bar{\mu}_b \perp \bar{x}_b
\]

Sorting on observables means, first, that variation in bureaucrat effects that cannot be captured by covariates must be orthogonal to non-bureaucrat shocks to test scores: \( \bar{\mu}_b \perp (v_b - \bar{\mu}_b) \). More subtly, we need that unobservable shocks to outcomes be independent of assignments: \((v_b - \bar{\mu}_b) \perp S_b \mid x_b, \bar{x}_b\).

<table>
<thead>
<tr>
<th>Violates Orthogonality Restriction</th>
<th>Doesn’t Violate Orthogonality Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>• High-VA bureaucrats are assigned to better districts; no controls for district quality</td>
<td>• High-VA bureaucrats are assigned to better districts, district quality is fixed, and district fixed effects are controlled for</td>
</tr>
<tr>
<td>• All bureaucrats have the same VA, but better-connected bureaucrats are assigned to better districts; no controls for district quality</td>
<td>• Bureaucrat quality increases linearly over time; time controls included</td>
</tr>
<tr>
<td>• Some states grow faster; since transfers are mostly within-state, bureaucrats assigned to some states will appear to get consistently good outcomes; state growth is not controlled for</td>
<td></td>
</tr>
</tbody>
</table>
We include controls that make the sorting on observables requirement more plausible. In our baseline model, we control for a quadratic time trend. In our second specification, we additionally include the following controls:

- **State-specific quadratic time trend**: A quadratic in months since start interacted with state indicator variables.

- **Lags**: The district’s mean outcome for each of the last three postings.

- **Cumulative mean outcome**: The district’s mean outcome over all previous postings.

In our final model, we also include district fixed effects. With district fixed effects, sorting on observables requires only that assignments be independent of transient shocks to districts, not to the persistent component of district quality. However, the large number of parameters this introduces presents challenges akin to those found when attempting to jointly estimate worker effects and firm effects, as in [Abowd et al.](1999).

One of the challenges created by controlling for district fixed effects is that it creates many more problems to estimate. This is a problem, at the least, for maximum likelihood. MLE estimates the variance of bureaucrat effects by decomposing effects into an unpredictable and a predictable component:

\[
\hat{\text{Var}}(\mu_i) = \sigma^2 + \text{Var}(\tilde{x}_i^T \hat{\lambda})
\]

The estimate of the predictable component is biased upwards from noisy estimates of \(\hat{\lambda}\), which describes the relationship between covariates and bureaucrat effects, and the bias is worse when \(\hat{\lambda}\) is high-dimensional, as in the case of district fixed effects. Bias-corrected maximum likelihood may not fix this issue when \(\hat{\lambda}\) is high dimensional, since the bias correction depends on an asymptotic approximation to the variance of \(\hat{\lambda}\). These issues are related to the “limited mobility bias,” described in [Andrews et al.](2008) and [Andrews et al.](2012), who show that the regression-based methods of [Abowd et al.](1999) have negatively biased estimates of the correlation between worker and firm effects and positively biased
estimates of the variance of worker effects and of firm effects. These issues are worse when there are few transfers of workers to firms or, in our data, of bureaucrats to districts.

Another challenge is that district fixed effects and bureaucrat effects are only identified within a “connected set.” A connected set is a set of districts and bureaucrats such that a path of bureaucrat transfers connects each district to each bureaucrat. This issue has been noted by Abowd et al. (1999). In a least-squares regression that includes district and bureaucrat dummies, one could remove this source of multicollinearity by dropping one bureaucrat fixed effect or one district fixed effect from each connected set. However, since we want to understand the variance of bureaucrat sets, results depend on which fixed effects we drop. Therefore, we produce results only within the largest connected set for each outcome, which contains 82% of observations, 81% of districts, and 65% of states (larger states are less likely to be outside the connected set).

Estimates using a placebo outcome (Table 3.4) show that specifications with district fixed effects estimated with MLE, bias-corrected MLE, or the modified-KS estimator are positively biased.

3.3.3 Estimators

We use several estimation strategies. Each of these estimators follows a two-stage estimation procedure. First, we estimate several population parameters: the variances of collector effects $\sigma^2_{\mu}$, posting effects specific to a collector-district pair $\sigma^2_{q}$, and month-district level shocks $\sigma^2_{r}$. Then, in order to estimate individual effects, we compute a preliminary estimate of each bureaucrat’s value-added as the residual from regressing the outcome of interest on district covariates. Finally, we use the parameters estimated in the first step to account for regression to the mean: we shrink the bureaucrat’s residual to make it a best linear predictor of the bureaucrat’s effect on the outcome.
**Estimator from Kane and Staiger (2008)**

The estimator from Kane and Staiger (2008) estimates parameters ($\eta$) by, first, residualizing away the effect of covariates then, second, exploiting the fact that the expected product of residuals in different classes (districts) with the same teacher (District Collector) is $\sigma^2_{\mu}$. This estimator assumes that $\lambda = 0$ — there is no correlation between bureaucrat effects and covariates — so $\widehat{\text{Var}}(\mu_j) = \hat{\sigma}^2_{\mu^*}$.

Let the set of districts bureaucrat $b$ serves in be $D(b)$, and call the mean outcomes and a residuals from a posting with bureaucrat $b$ and district $d$ be $\bar{y}_{bd}$ and $\bar{x}_{bd}$. In equations, this estimator is

$$\hat{b} = \arg \min_b \sum_{d,t} \left( y_{dt} - \bar{x}_{dt}^T b \right)^2$$

$$\hat{\sigma}^2_{\mu} = \frac{2}{\sum_b |D(b)| (|D(b)| - 1)} \sum_b \sum_{d,d'} \left( y_{bd} - \bar{x}_{bd}^T \hat{b} \right) \left( y_{bd'} - \bar{x}_{bd'}^T \hat{b} \right)$$

As Equation 3.3 illustrates, this estimator is inconsistent when $\lambda \neq 0$ due to omitted variable bias. When there is correlation between bureaucrat effects and covariates, this effects of $\lambda$ are instead attributed to $\hat{b}$, so that bureaucrat effects are attributed to covariates. For example, if bureaucrat quality is improving over time and we control for year, the Kane and Staiger estimator will assume that bureaucrat quality is constant over time and that outcomes have improved due to the increasing year. This negatively biases $\widehat{\text{Var}}(\mu_j)$.

**Modified KS**

The “modified-KS” estimator uses a modification inspired by [Chetty et al.] (2014a). It uses within-teacher (bureaucrat) variation to estimate $\hat{b}$, correcting the omitted variable bias problem that occurs when using the Kane and Staiger estimator. However, the modified-KS estimator can generate its own problems: Although consistent, it can be biased either up or down in finite samples, and since it does not exploit between-bureaucrat variation its estimate of $\hat{b}$ will be less precise, and perform especially poorly when there are errors in
variables.

Letting $\tilde{y}_{dt}$ and $\tilde{x}_{dt}$ be deviations from bureaucrat-level means, this estimator is given by

$$\hat{b} = \arg \min_b \sum_{d,t} \left( \tilde{y}_{dt} - \tilde{x}_{dt}^T b \right)^2$$

$$\hat{\sigma}_m^2 = \frac{2}{\sum_b |D(b)| (|D(b)| - 1)} \sum_b \sum_{d,d' \in D(b), d \neq d'} \left( \tilde{y}_{bd} - \tilde{x}_{bd}^T \hat{\beta} \right) \left( \tilde{y}_{bd'} - \tilde{x}_{bd'}^T \hat{\beta} \right).$$

**Maximum Quasi-Likelihood**

MLE can give more precise estimates because it can use information from collectors who only serve in one district, a large part of our sample. Although it will typically be biased upwards in finite samples, a bias correction is available. We report both raw and bias-corrected estimates.

A likelihood function is generated by assuming that all disturbances are normal:

$$y_{b} | x_{b}, \bar{x}_{b}, s_{b} \sim N \left( x_{b} \hat{\beta} + \bar{x}_{b}^T \lambda, \Sigma(\eta, s_{b}) \right)$$

Optimizing this likelihood function produces consistent estimates of all parameters even when the true DGP is not normal, making this a quasi-likelihood estimator.

But maximum (quasi-)likelihood has a drawback: its estimate of $\text{Var}(\mu_{j})$ is biased upwards. To see this intuitively, note that

$$\text{Var}(\mu_{b}) = \hat{\sigma}_m^2 + \text{Var} \left( \bar{x}_{b}^T \hat{\lambda} \right)$$

and consider that the second term of Equation (3.4) will be positive even when $\lambda = 0$. To correct this, we also report results using a bias correction from the second chapter of this dissertation.

**3.3.4 Randomization Inference**

In the presence of large finite-sample biases, standard errors and t-tests are an inappropriate way of measuring significance. Instead, we test significance by randomly re-labeling
bureaucrats, and deriving the distribution of $\widehat{\text{Var}}(\mu_j)$ for simulated bureaucrats. Simulations create datasets from a world in which bureaucrats have no effects on outcomes, so the distribution of the estimator on simulated data is the same as the distribution of the estimator under the null hypothesis that bureaucrats are identical. P-values are the fraction of $\widehat{\text{Var}}(\mu_j)$ in simulated datasets that are higher than our real estimates.

Simulations must be realistic: A key feature of our dataset is that transfers within districts in the same state are frequent, but transfers between states are not. Therefore, we simulate in a way that keeps the set of bureaucrats in any state at a given time invariant, approximately preserves the length and number of postings, and does not change the set of states a bureaucrat appears in. Within each state-month, if district has the same bureaucrat for two periods in a row in the real data, it will also not experience a change of bureaucrat in the simulated data. The bureaucrats and districts who change in the real data are randomly permuted within a state in the simulated data. Our simulations thus maintain the feature that it is relatively easy to compare bureaucrats serving in the same state around the same time, since they will often switch districts with each other, but that connections between bureaucrats in different states and different times are weak.

### 3.4 Results

Our results are sensitive to the choice of estimator and to the choice of covariates. Regardless, p-values tell a consistent story: our estimates cannot be differentiated from zero, and the large dispersion of simulation estimates suggests that bureaucrats would need to be a very large contributor to variance for their effects to be statistically significant. Estimates of $\widehat{\text{Var}}(\mu_j)$ are lowest with the Kane and Staiger estimator, consistent with the theoretical result that this estimator is asymptotically weakly negatively biased. Results from MLE without a bias correction are highest, consistent with the prediction that this estimator is positively biased in finite samples. Estimates from the modified-KS and bias-corrected MLE estimators, which we prefer, are typically in the middle. To validate these predictions, we use a placebo standard normal outcome, distributed $y_{dt} \sim \text{iid } N(0,1)$, to create a dataset
where \( \text{Var}(\mu_j) = 0 \). With the placebo outcome, the Kane and Staiger estimates are near zero and all other estimates are positive, suggesting the possibility of substantial finite-sample biases.

Table 3.2 shows estimates of the contribution of bureaucrats to variance in night light intensity, with a point estimate and p-value (from simulations) for each specification and each estimator. Figure 3.4 plots a histogram and empirical CDF of estimates from the second specification and the modified-Kane and Staiger estimator. The simplest specification, including only a quadratic time trend, gives unreasonably high answers. This could be the case because some states tend to have higher night light intensity and most bureaucrat transfers are within state, so variation in state light intensity will be mistakenly attributed to bureaucrats. The next specification controls for a state-specific quadratic trend, the mean outcome in each of the last three postings, and the cumulative mean outcome until the start of the posting, which hopefully accounts for the fact that different districts are on different growth trajectories. It gives smaller answers, but they are still large: our preferred estimators, Mod-KS and bias-corrected MLE, suggest that bureaucrats account for 11.17% to 14.69% of the variation in night light intensity, larger than the share of test score variance explained by bureaucrat quality. If bureaucrats account for 11% of the variance in night light intensity, a bureaucrat one standard deviation above average would increase night light intensity.
intensity by 0.33 standard deviations. Accounting for district fixed effects should make the sorting on observables requirement much more plausible, but district fixed effects appear to make the estimates unreasonably large. The large effects could be due to limited mobility bias or having too many covariates, as discussed above.

Figure 3.4 makes the high dispersion of the estimator clear. The fifth and ninety-fifth percentiles of simulated estimates of $\hat{\text{Var}}(\mu_b)$ are at 4% and 89%.

Table 3.3, showing the contribution of bureaucrats to variance in project completions, gives reasonably small estimates with the first two specifications but unreasonably large estimates with the last. Our preferred specifications and estimators suggest that bureaucrats account for 3.16% to 5.95% of the variation in project completions. If bureaucrats account for 1% of the variance, a bureaucrat one standard deviation above averages increases project completions by 0.1 standard deviations, which equals 0.05 projects per month, or 0.91
Table 3.3: Contribution of District Collectors to variation in project completions.

<table>
<thead>
<tr>
<th>specification</th>
<th>KS</th>
<th>Mod-KS</th>
<th>MLE</th>
<th>MLE Bias Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic Trend</td>
<td>Point</td>
<td>0.5%</td>
<td>0.7%</td>
<td>1.0%</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>[0.2%, 0.8%]</td>
<td>[0.4%, 1.1%]</td>
<td>[0.7%, 1.3%]</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.80</td>
<td>0.34</td>
<td>0.26</td>
</tr>
<tr>
<td>+ Lags</td>
<td>Point</td>
<td>-0.0%</td>
<td>3.2%</td>
<td>6.0%</td>
</tr>
<tr>
<td>State-Specific Trend, Cum. Mean</td>
<td>CI</td>
<td>[0%, 0.1%]</td>
<td>[2.6%, 3.7%]</td>
<td>[5.7%, 6.3%]</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.86</td>
<td>0.50</td>
<td>0.41</td>
</tr>
<tr>
<td>+ District FEs</td>
<td>Point</td>
<td>-0.0%</td>
<td>9.8%</td>
<td>15.8%</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>[0%, 0.1%]</td>
<td>[8.7%, 10.9%]</td>
<td>[15.4%, 16.1%]</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.78</td>
<td>0.12</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Table 3.4: Contribution of District Collectors to variation in placebo outcome.

<table>
<thead>
<tr>
<th>specification</th>
<th>KS</th>
<th>Mod-KS</th>
<th>MLE</th>
<th>MLE Bias Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic Trend</td>
<td>Point</td>
<td>0.17%</td>
<td>0.18%</td>
<td>0.03%</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>[0%, 0.61%]</td>
<td>[0%, 0.61%]</td>
<td>[0%, 0.19%]</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.29</td>
<td>0.30</td>
<td>0.32</td>
</tr>
<tr>
<td>+ Lags</td>
<td>Point</td>
<td>0.15%</td>
<td>6.40%</td>
<td>8.34%</td>
</tr>
<tr>
<td>State-Specific Trend, Cum. Mean</td>
<td>CI</td>
<td>[0%, 0.57%]</td>
<td>[5.39%, 7.42%]</td>
<td>[7.88%, 8.80%]</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.27</td>
<td>0.39</td>
<td>0.51</td>
</tr>
<tr>
<td>+ District FEs</td>
<td>Point</td>
<td>0.12%</td>
<td>5.32%</td>
<td>8.74%</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>[0%, 0.52%]</td>
<td>[4.48%, 6.17%]</td>
<td>[8.21%, 9.28%]</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.32</td>
<td>0.89</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Estimates using a placebo outcome, in Table 3.4, are much lower but are still positive, when they should be zero. Our preferred estimates, from the second specification and from the Mod-KS and bias-corrected MLE estimates, show that bureaucrats contribute 6.40% to
Figure 3.5: Distribution of simulation estimates for the contribution of District Collectors to variation in number of project completions. From the second specification and Modified-KS estimator.

8.26% of the variation in the placebo outcome, suggesting an upward bias in our estimates. Figure 3.6 plots the distribution of simulation estimates.

Together, these results are suggestive of bureaucrats contributing to variation in night light intensity and project completions but do not prove it. Estimates from real outcomes have higher point estimates and lower p-values than estimates from the placebo outcome, but still none of our estimates are significant at the 5% level.
Figure 3.6: Distribution of simulation estimates for the contribution of District Collectors to variation in the placebo outcome. From the second specification and Mod-KS estimator.

3.5 Conclusion

Value-added estimation in a setting with relatively few observations and high-dimensional covariates presents two econometric challenges. First, estimates of the variance of bureaucrat effects are biased upwards. It is possible, at great computational cost, to check whether bureaucrats vary significantly in their effects through randomization inference. Second, our estimates are high-variance; although accounting for 2% of variance in district-level light intensity is a very large effect economically, simulations suggest that the dispersion of the estimator (under the null hypothesis) is much larger.
References


Appendix A

Supplementary Figures for Chapter 1

Figure A.1: Magnitude of ELA teacher effects.

Notes: The top plot shows the variance of English teachers’ effects on English Language Arts scores, in the same year that the student has this teacher and 1, 2, 3, and 4 years after. The second and third plots show the variances of English teachers’ effects on math scores and attendance. Error bars plot a 95% credible interval based on 1000 Bayesian Bootstrap iterations.
### Table A.1: Magnitude of ELA teacher effects on test scores and attendance.

Notes: The standard deviation of English teacher effects on test scores and attendance, zero to four years out. This table displays the same information as Table A.1. 95% credible interval from 1000 Bayesian Bootstrap iterations in parentheses.

<table>
<thead>
<tr>
<th>Year</th>
<th>ELA score</th>
<th>Math score</th>
<th>Attendance Z-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.116</td>
<td>0.165</td>
<td>0.071</td>
</tr>
<tr>
<td>0</td>
<td>(0.112, 0.121)</td>
<td>(0.160, 0.170)</td>
<td>(0.069, 0.075)</td>
</tr>
<tr>
<td>1</td>
<td>0.104</td>
<td>0.135</td>
<td>0.072</td>
</tr>
<tr>
<td>2</td>
<td>0.119</td>
<td>0.144</td>
<td>0.078</td>
</tr>
<tr>
<td>3</td>
<td>0.117</td>
<td>0.140</td>
<td>0.079</td>
</tr>
<tr>
<td>4</td>
<td>0.099</td>
<td>0.139</td>
<td>0.080</td>
</tr>
<tr>
<td></td>
<td>(0.110, 0.125)</td>
<td>(0.140, 0.150)</td>
<td>(0.075, 0.083)</td>
</tr>
<tr>
<td></td>
<td>(0.113, 0.125)</td>
<td>(0.133, 0.149)</td>
<td>(0.075, 0.085)</td>
</tr>
<tr>
<td></td>
<td>(0.110, 0.125)</td>
<td>(0.119, 0.163)</td>
<td>(0.068, 0.095)</td>
</tr>
</tbody>
</table>

### Table A.2: Magnitude of ELA teacher effects on graduation.

Notes: The standard deviation of English teacher effects on four-year high school graduation. This table displays the same information as Table A.1. 95% credible interval in parentheses.

<table>
<thead>
<tr>
<th>Standard Deviation of Teacher Effect</th>
<th>Graduated, 4-year</th>
<th>Regents Diploma, 4-year</th>
<th>Advanced Regents Diploma, 4-year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.048</td>
<td>0.068</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>(0.046, 0.060)</td>
<td>(0.066, 0.077)</td>
<td>(0.059, 0.066)</td>
</tr>
</tbody>
</table>
Figure A.2: Fade-out of ELA teacher effects.

Notes: The best linear predictor coefficient of a teacher’s effect on a future outcome given her effect on a present outcome. Error bars plot a 95% credible interval based on 1000 Bayesian Bootstrap iterations.
<table>
<thead>
<tr>
<th>Year</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA score</td>
<td>0.160</td>
<td>0.173</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>(0.073, 0.252)</td>
<td>(0.113, 0.242)</td>
<td>(0.000, 0.000)</td>
<td>(0.000, 0.000)</td>
<td>(1.000, 1.000)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.071</td>
<td>-0.026</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>(-0.117, -0.026)</td>
<td>(-0.058, 0.008)</td>
<td>(0.000, 0.000)</td>
<td>(0.000, 0.000)</td>
<td>(1.000, 1.000)</td>
</tr>
<tr>
<td>Attendance Z-Score</td>
<td>0.095</td>
<td>0.070</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>(0.023, 0.167)</td>
<td>(0.016, 0.120)</td>
<td>(0.000, 0.000)</td>
<td>(0.000, 0.000)</td>
<td>(1.000, 1.000)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Graduated (4-year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA score</td>
<td>0.514</td>
<td>0.425</td>
<td>0.263</td>
<td>0.327</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>(0.457, 0.575)</td>
<td>(0.353, 0.503)</td>
<td>(0.172, 0.355)</td>
<td>(0.185, 0.481)</td>
<td>(-0.057, 0.097)</td>
</tr>
<tr>
<td>Math score</td>
<td>0.381</td>
<td>0.174</td>
<td>0.054</td>
<td>-0.071</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.348, 0.414)</td>
<td>(0.130, 0.215)</td>
<td>(0.002, 0.110)</td>
<td>(-0.172, 0.029)</td>
<td>(-0.034, 0.032)</td>
</tr>
<tr>
<td>Attendance Z-Score</td>
<td>0.477</td>
<td>0.335</td>
<td>0.229</td>
<td>0.153</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>(0.420, 0.535)</td>
<td>(0.270, 0.398)</td>
<td>(0.153, 0.300)</td>
<td>(0.048, 0.255)</td>
<td>(0.022, 0.214)</td>
</tr>
</tbody>
</table>

### Table A.3: Fade-out of ELA teacher effects.

*Notes:* Best linear predictor coefficients. 95% credible interval based on 1000 Bayesian Bootstrap iterations in parentheses.

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA score (4 years later)</td>
<td>0.270</td>
<td>0.085</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>(0.120, 0.520)</td>
<td>(0.014, 0.216)</td>
<td>(0.168, 0.593)</td>
</tr>
<tr>
<td>Math score (4 years later)</td>
<td>0.043</td>
<td>0.011</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>(0.009, 0.118)</td>
<td>(0.000, 0.057)</td>
<td>(0.018, 0.136)</td>
</tr>
<tr>
<td>Attendance Z-Score (4 years later)</td>
<td>0.008</td>
<td>0.018</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>(0.001, 0.036)</td>
<td>(0.002, 0.049)</td>
<td>(0.007, 0.061)</td>
</tr>
<tr>
<td>Graduated, 4-year</td>
<td>0.008</td>
<td>0.030</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>(0.000, 0.059)</td>
<td>(0.001, 0.085)</td>
<td>(0.007, 0.111)</td>
</tr>
<tr>
<td>Regents Diploma, 4-year</td>
<td>0.025</td>
<td>0.000</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>(0.002, 0.094)</td>
<td>(0.000, 0.014)</td>
<td>(0.005, 0.095)</td>
</tr>
<tr>
<td>Advanced Regents Diploma, 4-year</td>
<td>0.065</td>
<td>0.011</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>(0.036, 0.104)</td>
<td>(0.002, 0.027)</td>
<td>(0.039, 0.112)</td>
</tr>
</tbody>
</table>

### Table A.4: Goodness of proxy for ELA teachers.

*Notes:* Goodness of proxy for graduation and four-year-lead test scores and attendance, using same-year test scores and attendance. 95% credible set based on 1000 Bayesian Bootstrap iterations in parentheses.
Figure A.3: Magnitude of ELA teacher effects by grade.

Notes: The variance of ELA teachers’ effects on outcomes, or the diagonal of $\Sigma_\mu$, within each year. Error bars plot a 95% credible interval based on 1000 Bayesian Bootstrap iterations.

Table A.5: Standard deviation of ELA teacher effects on ELA scores: Robustness to choice of controls and estimator.
Figure A.4: Fade-out of ELA teacher effects by grade.

Notes: The best linear predictor coefficient for predicting a teacher’s effect on a future outcome given her effect on a present outcome and covariates, by grade. Error bars plot a 95% credible interval based on 1000 Bayesian Bootstrap iterations.

<table>
<thead>
<tr>
<th>Controls</th>
<th>Estimator</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Baseline</td>
<td>Moment-Matching</td>
<td>0.179</td>
</tr>
<tr>
<td>Baseline</td>
<td>MLE</td>
<td>0.191</td>
</tr>
<tr>
<td>Polynomial</td>
<td>Moment-Matching</td>
<td>0.187</td>
</tr>
<tr>
<td>Polynomial</td>
<td>MLE</td>
<td>0.200</td>
</tr>
<tr>
<td>All Lags</td>
<td>Moment-Matching</td>
<td>0.179</td>
</tr>
<tr>
<td>All Lags</td>
<td>MLE</td>
<td>0.194</td>
</tr>
</tbody>
</table>

Table A.6: Standard deviation of English teacher effects on math scores: Robustness to choice of controls and estimator.
Figure A.5: Eigenvalues of $\tilde{\Sigma}_{\mu}$ for English teachers.

Notes: Error bars show a 95% credible interval based on 1000 Bayesian Bootstrap iterations.

<table>
<thead>
<tr>
<th>Controls</th>
<th>Estimator</th>
<th>Year 0</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Moment-Matching</td>
<td>0.073</td>
<td>0.074</td>
<td>0.081</td>
<td>0.081</td>
<td>0.075</td>
</tr>
<tr>
<td>Baseline</td>
<td>MLE</td>
<td>0.078</td>
<td>0.084</td>
<td>0.090</td>
<td>0.090</td>
<td>0.096</td>
</tr>
<tr>
<td>Polynomial</td>
<td>Moment-Matching</td>
<td>0.077</td>
<td>0.077</td>
<td>0.085</td>
<td>0.093</td>
<td>0.118</td>
</tr>
<tr>
<td>Polynomial</td>
<td>MLE</td>
<td>0.087</td>
<td>0.086</td>
<td>0.092</td>
<td>0.097</td>
<td>0.125</td>
</tr>
<tr>
<td>All Lags</td>
<td>Moment-Matching</td>
<td>0.070</td>
<td>0.072</td>
<td>0.079</td>
<td>0.079</td>
<td>0.075</td>
</tr>
<tr>
<td>All Lags</td>
<td>MLE</td>
<td>0.083</td>
<td>0.082</td>
<td>0.089</td>
<td>0.090</td>
<td>0.100</td>
</tr>
</tbody>
</table>

Table A.7: Standard deviation of English teacher effects on attendance: Robustness to choice of controls and estimator.
Factors of ELA teacher effects

Figure A.6: Factors of $\bar{S}_m$ for ELA teachers.

Table A.8: Standard deviation of English teacher effects on graduation: Robustness to choice of controls and estimator.

<table>
<thead>
<tr>
<th>Controls</th>
<th>Estimator</th>
<th>Graduated</th>
<th>Regents Diploma</th>
<th>Advanced Regents Diploma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Moment-Matching</td>
<td>0.049</td>
<td>0.074</td>
<td>0.065</td>
</tr>
<tr>
<td>Baseline</td>
<td>MLE</td>
<td>0.051</td>
<td>0.076</td>
<td>0.065</td>
</tr>
<tr>
<td>Polynomial</td>
<td>Moment-Matching</td>
<td>0.059</td>
<td>0.073</td>
<td>0.077</td>
</tr>
<tr>
<td>Polynomial</td>
<td>MLE</td>
<td>0.059</td>
<td>0.072</td>
<td>0.062</td>
</tr>
<tr>
<td>All Lags</td>
<td>Moment-Matching</td>
<td>0.047</td>
<td>0.073</td>
<td>0.065</td>
</tr>
<tr>
<td>All Lags</td>
<td>MLE</td>
<td>0.052</td>
<td>0.078</td>
<td>0.065</td>
</tr>
</tbody>
</table>
Figure A.7: Robustness of ELA teacher effects to estimator.

Notes: Standard deviation of English teacher effects on present and future outcomes, for both the “moment-matching” estimator used above and maximum likelihood.
Figure A.8: Robustness of ELA teacher effects to choice of controls.

Notes: Standard deviation of English teacher effects on present and future outcomes, for the baseline controls used above, baseline controls plus all available lags, and for the controls from Chetty et al. (2014a), which include third-degree polynomials.
Appendix B

Software and Computation

The software used to run the computations in this dissertation is mainly in Python, especially using Numpy, Pandas, and Scipy. Code used to produce value-added estimates is at [github.com/esantorella/tva](https://github.com/esantorella/tva), supported by a suite of helper functions at [github.com/esantorella/hdfe](https://github.com/esantorella/hdfe) which provides meta-functions for working efficiently for grouped data; an algorithm to find connected sets in a bipartite graph; an algorithm for detecting and removing collinear columns; functions for creating dummy variables; and regression functionality.

As detailed in [2] I have used four algorithms to estimate value-added models. The function

```python
estimate_va
```

all four algorithms. It takes the following parameters:

- `a`  
- `data`  
  - all relevant data  
- `outcome`  
  - containing the outcome
• **teacher**
  containing teacher identifiers

• **covariates**
  as covariates, not including any categorical variables

• **categorical_controls**
  be expanded into dummy variables

• **moments_only**
  individual value-added scores or just their moments

• **jackknife**
  classroom-out estimates of individual value-added

• **method**
  
  - **Kane and Staiger (2008)**; ‘cfr’, for the “modified-KS” estimator with a modification used in **Chetty et al. (2014a)**; ‘fk’ for the estimator derived from **Fessler and Kasy (2017)**; or ‘mle’, for maximum likelihood estimation.

• **add_constant**
  be added to covariates

### B.1 Computation Time

Each value-added algorithm requires a least-squares regression, asymptotic runtime will be at least $O(NK^2)$ with $N$ observations and $K$ covariates. Since the Kane and Staiger algorithm does not involve teacher fixed effects in this regression, it will execute this regression fastest. The KS and CFR algorithms then work with residuals, running severeral simple $O(N)$ computations. (If the dataset is not already sorted by classroom id, sorting
it will take $O(N \log N)$. The FK and MLE algorithms both take much longer in practice, because they require numerical optimization.
Appendix C

Appendix to Chapter 2

C.1 Maximum Quasi-Likelihood Robustly Estimates Variances

If we assume the model of Section 2.3.1 in which data is drawn from some distribution \( D \) and we do not assume a functional form, then quasi-likelihood based on normality delivers consistent estimates of the parameters \( \eta = \left( \sigma^2 \mu, \sigma^2 \beta, \sigma^2 \lambda \right) \), with true value \( \eta^* = \left( \sigma^2 \mu, \sigma^2 \beta, \sigma^2 \lambda \right) \). Consider the normal model

\[
y_j | x_j, \bar{x}_j \sim N \left( x_j \beta + \bar{x}_j \lambda, \Sigma(\eta, s_j) \right),
\]

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

Lemma C.1.1.

\[
\eta^* = \eta_F \equiv \arg \max_{\eta} \mathbb{E}_D \log f \left( y_j | x_j, \bar{x}_j, s_j; \eta \right)
\]  \hspace{1cm} (C.1)

Proof.

\[
\mathbb{E}_D \log f \left( y_j | x_j, \bar{x}_j, s_j; \eta \right) = -\frac{1}{2} \log \det \Sigma(\eta, s_j)
\]

\[
- \frac{1}{2} \mathbb{E}_D \left[ \left( y_j - x_j \beta - \bar{x}_j \lambda \right)^T \Sigma(\eta, s_j)^{-1} \left( y_j - x_j \beta - \bar{x}_j \lambda \right) \right]
\]  \hspace{1cm} (C.2)

Since \( \beta \) corresponds to an unrestricted linear predictor, the values of \( \beta \) and \( \lambda \) that
maximize Equation C.2 do not depend on $\Sigma$, so
\[
\arg\max_{\beta, \lambda} E_D \log f(y_j|x_j, \bar{x}_j; \sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, \beta, \lambda) = \beta^*, \lambda^*
\]

After plugging in $\beta = \beta^*$ and $\lambda = \lambda^*$, we can rewrite Equation C.2 as
\[
E_D \log f(y_j|x_j, \bar{x}_j; \sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, \beta^*, \lambda^*)
\]
\[
= -\frac{1}{2} \log \det \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)
\]
\[
- \frac{1}{2} E_D \left[ \left( y_j - x_j\beta^* - \bar{x}_j^T \lambda^* \right)^T \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)^{-1} \left( y_j - x_j\beta^* - \bar{x}_j^T \lambda^* \right) | x_j, \bar{x}_j, s_j \right]
\]
\[
= -\frac{1}{2} \log \det \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)
\]
\[
- \frac{1}{2} \text{trace} \left( \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)^{-1} E_D \left[ \left( y_j - x_j\beta^* - \bar{x}_j^T \lambda^* \right)^T \left( y_j - x_j\beta^* - \bar{x}_j^T \lambda^* \right) \right] \right)
\]
\[
= -\frac{1}{2} \log \det \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j) - \frac{1}{2} \text{trace} \left( \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)^{-1} \Sigma(\eta^*, s_j) \right)
\]
\[
= -\frac{1}{2} \log \det \Sigma(\eta, s_j) - \frac{1}{2} \text{trace} \left( \Sigma(\eta, s_j)^{-1} \Sigma(\eta^*, s_j) \right)
\]

Therefore,
\[
\arg\max_{\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon} E_D \log f(y_j|x_j, \bar{x}_j; \sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, \beta, \lambda)
\]
\[
= \arg\max_{\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon} -\log \det \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j) - \text{trace} \left( \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)^{-1} \Sigma(\eta^*, s_j) \right)
\]
\[
= \arg\max_{\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon} \log \det \left( \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)^{-1} \Sigma(\eta^*, s_j) \right) - \text{trace} \left( \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)^{-1} \Sigma(\eta^*, s_j) \right)
\]

Let $\Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)^{1/2}$ be the symmetric, positive definite square root of the symmetric, positive definite matrix $\Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, x_j)$, and let $\{e_i\}$ be the eigenvalues of $\Sigma(\eta, s_j)^{1/2} \Sigma(\sigma^2_\mu, \sigma^2_\theta, \sigma^2_\epsilon, s_j)^{-1} \Sigma(\eta, s_j)^{1/2}$. Since that matrix is positive definite, all of its eigenvalues are positive.
\[
\log \det \left( \Sigma(\sigma^2_{\mu}, \sigma^2_{\theta}, \sigma^2_{\epsilon}, s_j)^{-1} \Sigma(\eta^*, s_j) \right) - \text{trace} \left( \Sigma(\sigma^2_{\mu}, \sigma^2_{\theta}, \sigma^2_{\epsilon}, s_j)^{-1} \Sigma(\eta^*, s_j) \right)
= \log \det \left( \Sigma(\eta^*, s_j)^{1/2} \Sigma(\sigma^2_{\mu}, \sigma^2_{\theta}, \sigma^2_{\epsilon}, s_j)^{-1} \Sigma(\eta^*, s_j)^{1/2} \right) - \text{trace} \left( \Sigma(\eta^*, s_j)^{1/2} \Sigma(\sigma^2_{\mu}, \sigma^2_{\theta}, \sigma^2_{\epsilon}, s_j)^{-1} \Sigma(\eta^*, s_j)^{1/2} \right)
= \log \prod_i e_i - \sum_i e_i
= \sum_i (\log(e_i) - e_i)
\]

Equation (C.3) is maximized when all \(e_i = 1\), which occurs when \(\Sigma(\sigma^2_{\mu}, \sigma^2_{\theta}, \sigma^2_{\epsilon}, s_j) = \Sigma(\eta^*, s_j)\). As long as the teacher teaches multiple classes and at least one class has multiple students, the only value that solves this equation is \(\eta_F = \eta^*\).

C.2 Closed-Form Likelihood and Intuitive Parameter Estimates

This section derives a closed-form solution for the likelihood. Subsection C.2.1 translates Equation (C.1), which is in terms of a determinant and an inverse of \(\Sigma\), into an equation that contains integrals but no determinant or inverse. Subsection C.2.2 solves these integrals to give a tractable formula for the likelihood.

C.2.1 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 \(y_j\). For each classroom \(c\), define \(\ell_c\), 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:
\[ \mu_j \sim N \left( \bar{x}_j^T \lambda, \sigma^2_{\mu} \right) \]
\[ \theta_c \sim N(0, \sigma^2_{\theta}) \]
\[ \varepsilon_j \sim N \left( x_j^T \beta, I \sigma^2_{\varepsilon} \right) \]  
(C.5)

Stack the \( \theta_c \) corresponding to each classroom into a vector \( \Theta_j \). The covariance of \( \Theta_j \) is block diagonal, with diagonal blocks corresponding to each classroom:

\[
\Theta_j = \begin{pmatrix}
\theta_1 \ell_1 \\
\theta_2 \ell_2 \\
\vdots \\
\theta_C \ell_C
\end{pmatrix}, \quad \text{Var}(\Theta_j) = \sigma^2_{\theta} B \\
B = \begin{pmatrix}
\ell_1 \ell_1^T & 0 & 0 & 0 \\
0 & \ell_2 \ell_2^T & 0 & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & 0 & \ell_C \ell_C^T
\end{pmatrix}.
\]

Summing the random variables of Equation C.5 gives a new random variable that has the same distribution as \( y_j \):

\[
\mu_j + \Theta_j + \varepsilon_j \sim N \left( x_j^T \beta + \bar{x}_j^T \lambda, \ell \ell^T \sigma^2_{\mu} + B \sigma^2_{\theta} + I \sigma^2_{\varepsilon} \right) \\
\ell \ell^T \sigma^2_{\mu} + B \sigma^2_{\theta} + I \sigma^2_{\varepsilon} = \Sigma_j(\eta) \\
\mu_j + \Theta_j + \varepsilon_j \overset{D}{=} y_j
\]

Intuitively, \( \mu_j \) affects all students taught by the same teacher, each \( \theta_c \) affects each student in classroom \( c \) and is independent from all other \( \theta_k \), and each component of \( \varepsilon_j \) 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:

\[
f_{y_j}(y_j|\eta) = f_{\mu_j+\Theta_j+\varepsilon_j}(y_j|\eta) \\
= \int_{\mu} f(\mu) f_{\Theta_j+\varepsilon_j}(y_j - \mu) \, d\mu \\
= \int_{\mu} \phi \left( \mu; \sigma^2_{\mu} \right) f_{\Theta_j+\varepsilon_j}(y_j - \mu) \, d\mu \\ 
(C.6)
\]
Since the covariance matrix of $\Theta_j + \varepsilon_j$ is block diagonal, we can write its probability density function as a product over the blocks, which correspond to classes:

$$f(\Theta_j + \varepsilon_j) = \Pi_c f(\ell_c \theta_c + \varepsilon_c)$$

$$\ell_c \theta_c + \varepsilon_c \sim N \left( x_i^T \beta, \ell_c \ell_c^T \sigma^2_\theta + I \sigma^2_\varepsilon \right)$$

$$f_{\ell_c \theta_c + \varepsilon_c}(y_c - \mu) = \int_\theta f(\theta) f_{\varepsilon_c}(y_c - \mu - \theta) \, d\theta$$

$$= \int_\theta \phi(\theta; \sigma^2_\theta) \Pi_{i \in I(c)} \phi \left( y_i - x_i^T \beta - \mu - \theta; \sigma^2_\varepsilon \right) \, d\theta$$

$$f_{\Theta_j + \varepsilon_j}(y_j - \mu) = \Pi_c \int_\theta \phi(\theta; \sigma^2_\theta) \Pi_{i \in I(c)} \phi \left( y_i - x_i^T \beta - \mu - \theta; \sigma^2_\varepsilon \right) \, d\theta$$

Plugging Equation C.7 into Equation C.6 we get a complete formula for the likelihood:

$$f(y_j | x_j, \tilde{x}_j, \tilde{s}_j; \theta, \beta, \lambda, a)$$

$$= \int_\mu \phi \left( \mu - \tilde{x}_j^T \lambda; \sigma^2_\mu \right) \Pi_c \left( \int_\theta \phi(\theta; \sigma^2_\theta) \Pi_{i \in I(c)} \phi \left( y_i - x_i^T \beta - \mu - \theta; \sigma^2_\varepsilon \right) \, d\theta \right) \, d\mu$$

C.2.2 Solving the Integrals

This section derives a closed-form solution for the likelihood using Equation C.9 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 $x$:

$$\bar{y}_c \equiv \frac{1}{|I(c)|} \sum_{i \in I(c)} y_i \quad (C.10)$$

$$\tilde{y}_i \equiv y_i - \bar{y}_{c(i)} \quad (C.11)$$

$$\bar{x}_c \equiv \bar{x}_c \beta - \mu_{j(c)}$$

$$\sigma^2_\mu \equiv \text{Var} \left( \bar{y}_c - \bar{x}_c \beta - \mu_{j(c)} \right) = \sigma^2_\theta + \sigma^2_\varepsilon / n_c. \quad (C.13)$$
Define precision-weighted teacher-level means and within-teacher demeaned values for $y$, and analogues for $x$:

\[ \bar{y}_j = \frac{\sum_{c:j(c)=j} h_c \bar{y}_c}{\sum_{c:j(c)=j} h_c} \]
\[ \bar{y}_c = \bar{y}_c - \bar{y}_{j(c)} \]  \hspace{1cm} (C.14)

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

\[ \Pi_{i \in I(c)} \phi \left( y_i; \sigma^2_c \right) \propto \sigma^{2-n_c}_c \phi \left( \sqrt{\sum_i \bar{y}_i^2}, \sigma^2 \right) \phi \left( \bar{y}_c, \sigma^2_c / n_c \right). \]  \hspace{1cm} (C.15)

Equation (C.15) implies

\[ \Pi_{i \in I(c)} \phi \left( y_i - x_i^T \beta - \mu - \theta; \sigma_c \right) \propto \sigma^{2-n_c}_c \phi \left( \sqrt{\sum_{i \in I(c)} \left( \bar{y}_i - \bar{x}_i^T \beta \right)^2}, \sigma^2_c \right) \phi \left( \bar{y}_c - \bar{x}_c^T \beta - \mu - \theta; \sigma^2_c / n_c \right), \]

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

\[ \int_{\mu} \phi(\mu - x_1; \sigma_1) \phi(\mu - x_2; \sigma_2) \, d\mu = \phi \left( x_1 - x_2; \sqrt{\sigma^2_1 + \sigma^2_2} \right). \]  \hspace{1cm} (C.16)

Applying Equation (C.16),

\[ \int_{\beta} \phi(\theta; \sigma_0) \Pi_{i} \phi \left( y_i - x_i^T \beta - \mu - \theta; \sigma_c \right) \, d\theta = \sigma^{2-n_c}_c \phi \left( \sum_i \left( \bar{y}_i - \bar{x}_i^T \beta \right)^2, \sigma^2_c \right) \int_{\beta} \phi(\theta; \sigma_0) \phi \left( \bar{y}_c - \bar{x}_c^T \beta - \mu - \theta; \sigma^2_c / n_c \right) \, d\theta \]
\[ = \sigma^{2-n_c}_c \phi \left( \sum_i \left( \bar{y}_i - \bar{x}_i^T \beta \right)^2, \sigma^2_c \right) \phi \left( \bar{y}_c - \bar{x}_c^T \beta - \mu; \sigma^2_0 + \sigma^2_c / n_c \right) \]
\[ = \sigma^{2-n_c}_c \phi \left( \sum_i \left( \bar{y}_i - \bar{x}_i^T \beta \right)^2, \sigma^2_c \right) \phi \left( \bar{y}_c - \bar{x}_c^T \beta - \mu; 1 / h_c \right). \]
Note the product of \( n \) normal densities with different means and variances:

\[
\Pi_c \phi(\mu_c; \sigma_c) = \sqrt{\frac{1}{\sum_c 1/\sigma_c^2}} \phi\left( \frac{1}{\sum_c 1/\sigma_c^2} \right) \phi\left( \mu_c \frac{1}{\sum_c 1/\sigma_c^2} \right)
\]

\[\equiv \sqrt{\frac{1}{\sum_c h_c}} \phi\left( \frac{\mu_c}{\sum_c h_c} \right) \phi\left( \mu_c \frac{1}{\sum_c h_c} \right) \quad (C.17)\]

Applying Equation \( C.17 \),

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

Therefore,

\[
\Pi_c \int \phi(\theta; \sigma_c) \Pi_i \phi(y_i - x_i \beta - \mu - \theta; \sigma_c) \ d\theta = \sigma_{c_0}^{1+N_{\text{classes}} - N_{\text{students}}} \sqrt{\frac{\Pi_c h_c}{\sum_c h_c}} \exp\left( -\frac{1}{2} \sum_c \left( \tilde{y}_c - \tilde{x}_c \beta \right)^2 h_c \right)
\]

\[\times \phi\left( \sqrt{\sum_i (y_i - x_i \beta)^2; \sigma^2_c} \right) \phi\left( \frac{\tilde{y}_j - \tilde{x}_j \beta - \mu, 1}{\sum_c h_c} \right)
\]

Applying Equation \( C.16 \) again,

\[
\int \phi\left( \mu - \tilde{x}_j \lambda; \sigma^2_{\mu} \right) \phi\left( \frac{\tilde{y}_j - \tilde{x}_j \beta - \mu, \frac{1}{\sum_c h_c}}{\mu} \right) \ d\mu = \phi\left( \frac{\tilde{y}_j - \tilde{x}_j (\beta + \lambda)}{\sigma^2_{\mu} + \frac{1}{\sum_c h_c}} \right)
\]

Equation \( C.9 \) finally reduces to

\[
f\left( y | x, s, \sigma^2_{\mu}, \sigma^2_{\beta}, \sigma^2_{\lambda}, \beta, \lambda \right) = \sigma_c^{1+N_{\text{classes}} - N_{\text{students}}} \sqrt{\frac{\Pi_c h_c}{\sum_c h_c}} \exp\left( -\frac{1}{2} \sum_c \left( \tilde{y}_c - \tilde{x}_c \beta \right)^2 h_c \right)
\]

\[\times \phi\left( \sqrt{\sum_i (y_i - x_i \beta)^2; \sigma^2_c} \right) \phi\left( \frac{\tilde{y}_j - \tilde{x}_j (\beta + \lambda)}{\sigma^2_{\mu} + \frac{1}{\sum_c h_c}} \right)
\]

The log-likelihood is
\[
\log f (y_j | x_j, \bar{x}_j, s; \theta, \beta, \lambda) \\
= (1 + N_{\text{classes}} - N_{\text{students}}) \log \sigma_e + \frac{1}{2} \sum_c \log (h_c) - \frac{1}{2} \log \left( \sum_c h_c \right) - \frac{1}{2} \sum_c \left( \bar{y}_c - \bar{x}_c^T \beta \right)^2 h_c \\
+ \log \phi \left( \sqrt{\sum_i (\bar{y}_i - \bar{x}_i^T \beta)^2 \sigma_e^2} \right) + \log \phi \left( \bar{y}_j - \bar{x}_j^T (\beta + \lambda), \sigma_e^2 + \frac{1}{\sum_c h_c} \right) \\
= (1 + N_{\text{classes}} - N_{\text{students}}) \log \sigma_e + \frac{1}{2} \sum_c \log (h_c) - \frac{1}{2} \log \left( \sum_c h_c \right) - \frac{1}{2} \sum_c \left( \bar{y}_c - \bar{x}_c^T \beta \right)^2 h_c \\
- \log \sigma_e - \frac{1}{2 \sigma_e^2} \sum_i (\bar{y}_i - \bar{x}_i^T \beta)^2 - \frac{1}{2} \log \left( \sigma^2 + \frac{1}{\sum_c h_c} \right) - \frac{1}{2} \left( \frac{1}{\sigma_e^2 + \frac{1}{\sum_c h_c}} \right) \left( \bar{y}_j - \bar{x}_j^T (\beta + \lambda) \right)^2 \\
= (N_{\text{classes}} - N_{\text{students}}) \log \sigma_e + \frac{1}{2} \sum_c \log (h_c) - \frac{1}{2} \log \left( \sum_c h_c \right) - \frac{1}{2} \sum_c \left( \bar{y}_c - \bar{x}_c^T \beta \right)^2 h_c \\
- \frac{1}{2 \sigma_e^2} \sum_i (\bar{y}_i - \bar{x}_i^T \beta)^2 - \frac{1}{2} \sum_j \log \left( \sigma^2 + \frac{1}{\sum_{c \in C(j)} h_c} \right) - \sum_j \frac{1}{2} \left( \frac{1}{\sigma_e^2 + \frac{1}{\sum_{c \in C(j)} h_c}} \right) \left( \bar{y}_j - \bar{x}_j^T (\beta + \lambda) \right)^2.
\]

The log-likelihood for all teachers is

\[
\sum_j \log f (y_j | x_j, \bar{x}_j, s; \sigma^2, \sigma^2, \sigma^2, \beta, \lambda) \\
= (N_{\text{classes}} - N_{\text{students}}) \log \sigma_e + \frac{1}{2} \sum_c \log h_c - \frac{1}{2} \sum_j \log \left( \sum_{c \in C(j)} h_c \right) - \frac{1}{2} \sum_c \left( \bar{y}_c - \bar{x}_c^T \beta \right)^2 h_c \\
- \frac{1}{2 \sigma_e^2} \sum_i (\bar{y}_i - \bar{x}_i^T \beta)^2 - \frac{1}{2} \sum_j \log \left( \sigma^2 + \frac{1}{\sum_{c \in C(j)} h_c} \right) - \sum_j \frac{1}{2} \left( \frac{1}{\sigma_e^2 + \frac{1}{\sum_{c \in C(j)} h_c}} \right) \left( \bar{y}_j - \bar{x}_j^T (\beta + \lambda) \right)^2.
\]

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

\[
f (y_j | x_j, s; \eta) = g (\eta) \int \phi \left( \mu - \bar{x}_j^T \lambda; \sigma^2 \right) \phi \left( \bar{y}_j - \bar{x}_j^T \beta - \mu; \frac{1}{\sum_c h_c} \right) \, d\mu \quad (C.18)
\]
C.3 Maximum Likelihood Bias Correction

Proof of Equation 2.6:

\[
\mathbb{E} \left[ \text{Var} \left( \hat{\mathbf{x}}^T \hat{\lambda} | \hat{\lambda} \right) \right] - \text{Var} \left( \hat{\mathbf{x}}^T \lambda \right) = \text{Var} \left( \hat{\mathbf{x}}^T \hat{\lambda} \right) - \text{Var} \left( \mathbb{E} \left[ \hat{\mathbf{x}}^T \hat{\lambda} | \hat{\lambda} \right] \right) - \text{Var} \left( \hat{\mathbf{x}}^T \lambda \right)
\]

\[
= \text{Var} \left( \hat{\mathbf{x}}^T \hat{\lambda} \right) - \text{Var} \left( \mathbb{E} \left[ \hat{\mathbf{x}}^T \hat{\lambda} \right] \right) - \text{Var} \left( \hat{\mathbf{x}}^T \lambda \right)
\]

\[
= \mathbb{E} \left[ \text{Var} \left( \hat{\mathbf{x}}^T \hat{\mathbf{x}} \right) \right] + \text{Var} \left( \mathbb{E} \left[ \hat{\mathbf{x}}^T \hat{\lambda} \right] \right) - \mathbb{E} \left[ \hat{\mathbf{x}}^T \right] \text{Cov}(\hat{\lambda}) \mathbb{E} [\hat{\mathbf{x}}] - \text{Var} \left( \hat{\mathbf{x}}^T \lambda \right)
\]

\[
= \mathbb{E} \left[ \text{Cov} \left( \hat{\mathbf{x}} \right) \hat{\lambda} \right] - \mathbb{E} \left[ \hat{\mathbf{x}}^T \right] \text{Cov}(\hat{\lambda}) \mathbb{E} [\hat{\mathbf{x}}]
\]

\[
= \mathbb{E} \left[ (\hat{\mathbf{x}} - \mathbb{E} \hat{\mathbf{x}})^T \text{Cov} (\hat{\lambda}) (\hat{\mathbf{x}} - \mathbb{E} \hat{\mathbf{x}}) \right].
\]

(C.19)

C.4 Optimization, Gradient

This equation can easily be optimized numerically. The software package available at http://www.github.com/esantorella/tva iterates between estimating \( \hat{\beta} \) and \( \hat{\lambda} \), which have closed-form solutions in terms of other parameters, and estimating \( \sigma^2_\mu \), \( \sigma^2_\theta \), and \( \sigma^2_\epsilon \) using L-BFGS.

C.4.1 Gradient

For compactness, let \( \eta_j \equiv \frac{1}{\sum_{c \in C(j)} \eta_c} \).
\[
\frac{\partial LL}{\partial \sigma^2_{\mu}} = \frac{1}{2} \sum_j \left( \frac{1}{\sigma^2_{\mu} + \eta_j} \right)^2 \left( (\bar{y}_j - \bar{x}_j^T \beta + \lambda)^2 - (\sigma^2_{\mu} + \eta_j) \right)
\]
\[
\frac{\partial LL}{\partial \sigma^2_{\theta}} = \frac{1}{2} \sum_c \frac{\partial h_c}{\partial \sigma^2_{\theta}} \left( \frac{1}{\sigma^2_{\mu} + \eta_c} - \eta_j \right)^2 \left( (\bar{y}_c - \bar{x}_c^T \beta)^2 \right)
\]
\[
\quad + \frac{1}{2} \sum_j \left( \sum_{c \in C(j)} \frac{\partial h_c}{\partial \sigma^2_{\theta}} \left( \frac{1}{\sigma^2_{\mu} + 1/\eta_j} - \left( \frac{\eta_j}{\sigma^2_{\mu} + \eta_j} \right)^2 \right) \left( \bar{y}_j - \bar{x}_j^T \beta + \lambda \right)^2 \right)
\]
\[
\quad - \sum_j \frac{1}{\sigma^2_{\mu} + \sum_{c \in C(j)} \frac{1}{\sigma^2_{\theta}} \left( \bar{y}_j - \bar{x}_j^T \beta + \lambda \right)^2 \left( \bar{y}_j - \bar{x}_j^T \beta + \lambda \right) h_c^2 - \sum_{c \in C(j)} h_c^2 (\bar{y}_c - \bar{x}_c^T \beta + \lambda) \right)
\]
\[
\frac{\partial LL}{\partial \sigma^2_{\epsilon}} = \frac{1}{2} \sum_c \frac{\partial h_c}{\partial \sigma^2_{\epsilon}} \left( \frac{1}{\sigma^2_{\mu} - (\bar{y}_c - \bar{x}_c^T \beta)^2} \right)
\]
\[
\quad + \frac{1}{2} \sum_j \left( \sum_{c \in C(j)} \frac{\partial h_c}{\partial \sigma^2_{\epsilon}} \left( \frac{1}{\sigma^2_{\mu} + 1/\eta_j} - \left( \frac{\eta_j}{\sigma^2_{\mu} + \eta_j} \right)^2 \right) \left( \bar{y}_j - \bar{x}_j^T \beta + \lambda \right)^2 \right)
\]
\[
\quad - \sum_j \frac{1}{\sigma^2_{\mu} + \sum_{c \in C(j)} \frac{1}{\sigma^2_{\epsilon}} \left( \bar{y}_j - \bar{x}_j^T \beta + \lambda \right)^2 \left( \bar{y}_j - \bar{x}_j^T \beta + \lambda \right) h_c^2 - \sum_{c \in C(j)} h_c^2 (\bar{y}_c - \bar{x}_c^T \beta + \lambda) \right)
\]
\[
\quad - \frac{1}{2} \frac{N_{\text{students}} - N_{\text{classes}}}{\sigma^2_{\epsilon}} + \frac{1}{\sigma^2_{\epsilon}} \sum_i (\bar{y}_i - \bar{x}_i^T \beta)^2
\]
\[
\frac{\partial LL}{\partial \lambda} = \sum_j \frac{1}{\sigma^2_{\mu} + \eta_j} \left( \bar{y}_j - \bar{x}_j^T \beta + \lambda \right) \bar{x}_j
\]
\[
\frac{\partial LL}{\partial \beta} = \sum_c (\bar{y}_c - \bar{x}_c^T \beta) \bar{x}_c h_c + \frac{1}{\sigma^2_{\mu}} \sum_i (\bar{y}_i - \bar{x}_i^T \beta) \bar{x}_i + \sum_j \frac{1}{\sigma^2_{\mu} + \eta_j} (\bar{y}_j - \bar{x}_j^T \beta + \lambda) \bar{x}_j
C.5 Bounding the Asymptotic Bias in the Kane and Staiger Procedure

We know that

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

\[= \text{Var}(\mu_j) - 2 \left( \sum_i \mu_{j(i)} x_i^T \right) \left( \sum_i x_i x_i^T \right)^{-1} \left( \sum_j |C(j)| |C(j)| - 1 \sum_{c, c' \in C(j)} \bar{x}_c \bar{x}_{c'} \right) \left( \sum_i x_i x_i^T \right)^{-1} \left( \sum_i \mu_{j(i)} x_i \right)
\]

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

\[
\text{Bias(\hat{\text{Var}}(\mu_j))} = -2 \left( \frac{1}{\text{N_{students}}} \sum_i x_i x_i^T \right)^{-1} \left( \frac{1}{\text{N_{students}}} \sum_i x_i x_i^T \right) \left( \frac{1}{\text{N_{teachers}}} \sum_j \mu_{j(i)} \bar{x}_j \right) \\
+ \left( \frac{1}{\text{N_{teachers}}} \sum_i x_i x_i^T \right)^{-1} \left( \frac{1}{\text{N_{classroompairs}}} \sum_{c, c' \in C(j)} \bar{x}_c \bar{x}_{c'} \right) \left( \frac{1}{\text{N_{students}}} \sum_i x_i x_i^T \right)^{-1} \left( \frac{1}{\text{N_{teachers}}} \sum_j \mu_{j(i)} \bar{x}_j \right)
\]

We want to show that

\[-b^T \left( \frac{1}{\text{N_{students}}} \sum_i x_i x_i^T \right) b \leq b^T \left( \frac{1}{\text{N_{students}}} \sum_i x_i x_i^T \right) b \quad \forall b \in \mathcal{R}^k.
\]

(C.22)
Note that

\[
\frac{1}{N_{\text{students}}} \sum_i x_i x_i^T = \frac{1}{N_{\text{students}}} \sum_i \left( \tilde{x}_j(i) + \tilde{x}_{c(i)} + \tilde{x}_i \right) \left( \tilde{x}_j(i) + \tilde{x}_{c(i)} + \tilde{x}_i \right)^T
\]

\[
= \frac{1}{N_{\text{teachers}}} \sum_j \tilde{x}_j \tilde{x}_j^T + \frac{1}{N_{\text{classrooms}}} \sum_c \tilde{x}_c \tilde{x}_c^T + \frac{1}{N_{\text{students}}} \sum_i \tilde{x}_i \tilde{x}_i
\]

\[
\frac{1}{N_{\text{class pairs}}} \sum_j \sum_{c', \in C(j)} \tilde{x}_c \tilde{x}_{c'}^T = \frac{1}{N_{\text{class pairs}}} \sum_j \sum_{c', \in C(j)} \left( \tilde{x}_j + \tilde{x}_c \right) \left( \tilde{x}_j + \tilde{x}_{c'} \right)
\]

\[
= \frac{1}{N_{\text{teachers}}} \tilde{x}_j \tilde{x}_j^T + \frac{1}{N_{\text{classrooms}}} \sum_c \frac{1}{|C(j)|} \sum_j \sum_{c', \in C(j)} \tilde{x}_c \tilde{x}_{c'}^T
\]

\[
= \frac{1}{N_{\text{teachers}}} \tilde{x}_j \tilde{x}_j^T - \frac{1}{N_{\text{classrooms}}} \sum_c \frac{1}{|C(j)|} \sum_j \sum_{c', \in C(j)} \tilde{x}_c \tilde{x}_{c'}^T
\]

Both the right hand side of both equations in Equation C.23 are sums of positive definite matrices, proving Equation C.22. Therefore, we know that

\[
\left| \left( \frac{\sum_j \mu_j \tilde{x}_j^T}{N_{\text{teachers}}} \right) \left( \frac{\sum_i x_i x_i^T}{N_{\text{students}}} \right)^{-1} \left( \frac{\sum_{c', \in C(j)} \tilde{x}_c \tilde{x}_{c'}^T}{N_{\text{class pairs}}} \right) \left( \frac{\sum_i x_i x_i^T}{N_{\text{students}}} \right)^{-1} \left( \frac{\sum_j \mu_j \tilde{x}_j}{N_{\text{teachers}}} \right) \right| 
\]

\[
\leq \left| \left( \frac{\sum_j \mu_j \tilde{x}_j^T}{N_{\text{teachers}}} \right) \left( \frac{\sum_i x_i x_i^T}{N_{\text{students}}} \right)^{-1} \left( \frac{\sum_j \mu_j \tilde{x}_j}{N_{\text{teachers}}} \right) \right|
\]

Plugging this into Equation C.20 we find that

\[
\text{Bias(Var}(\mu_j)) \leq - \left( \frac{\sum_j \mu_j \tilde{x}_j^T}{N_{\text{teachers}}} \right) \left( \frac{\sum_i x_i x_i^T}{N_{\text{students}}} \right)^{-1} \left( \frac{\sum_j \mu_j \tilde{x}_j}{N_{\text{teachers}}} \right)
\]

\[
\text{Bias(Var}(\mu_j)) \geq -3 \left( \frac{\sum_j \mu_j \tilde{x}_j^T}{N_{\text{teachers}}} \right) \left( \frac{\sum_i x_i x_i^T}{N_{\text{students}}} \right)^{-1} \left( \frac{\sum_j \mu_j \tilde{x}_j}{N_{\text{teachers}}} \right).
\]