The Effects of Teacher Quality on Adult Criminal Justice Contact

Evan K. Rose, Jonathan Schellenberg, and Yotam Shem-Tov*

December 2022

Abstract

This paper estimates teachers' impacts on their students' future criminal justice contact (CJC). Using a unique data set linking the universe of North Carolina public school data to administrative arrest records, we find a standard deviation of teacher effects on students' future arrests of 2.7 percentage points (11% of the sample mean). Teachers' effects on CJC are orthogonal to their effects on academic achievement, implying assignment to a high test score value-added teacher does not reduce future CJC. However, teachers who reduce suspensions and improve attendance substantially reduce future arrests. Similar patterns emerge when allowing teacher impacts to vary by student sex, race, socio-economic status, and school. The results suggest that the development of non-cognitive skills is central to the returns to education for crime and highlight an important dimension of teachers' social value missed by test score-based quality metrics.

*Rose: Assistant Professor, University of Chicago and NBER; ekrose@uchicago.edu. Schellenberg: Economist at Amazon Web Services; jschellenberg@econ.berkeley.edu. Shem-Tov: Assistant Professor, University of California, Los Angeles; shemtov@econ.ucla.edu. We are extremely grateful to Christopher Walters, David Card, Jesse Rothstein, and Patrick Kline, who provided invaluable support and advice. We thank the North Carolina Education Research Data Center for assistance in the construction of our data set and to Justin McCrary and the Spencer Foundation, who helped fund this project. Thanks also go to Natalie Bau, Gordon Dahl, Zarek Brot-Goldberg, Alessandra Fenizia, Kevin Todd, Nicholas Y. Li, Conrad Miller, Hilary Hoynes, Jonathan Holmes, Juliana Londoño-Vélez, Jennifer Kwok, Julien LaFortune, John Loeser, Justin McCrary, Enrico Moretti, Derek Neal, Elena Pastorino, Avner Shlain, and Isaac Sorkin for their helpful comments and suggestions. A previous version of this paper was entitled "The Effects of Teacher Quality on Criminal Behavior."

1 Introduction

Human capital is an important driver of criminal justice contact (CJC). Increasing schooling and school quality decreases the likelihood of future arrest, conviction, and incarceration (Lochner and Moretti, 2004; Deming, 2009; Heckman et al., 2010a; Lochner, 2011; Cook and Kang, 2016; Bell, Costa and Machin, 2018). Less is known, however, about how teacher quality affects CJC, despite teachers' central role in providing education and substantial research on their effects on test scores (Rivkin, Hanushek and Kain, 2005; Kane and Staiger, 2008; Chetty, Friedman and Rockoff, 2014*a*; Bacher-Hicks, Kane and Staiger, 2014*a*; Bau and Das, 2020) and long-run academic and economic outcomes (Chetty, Friedman and Rockoff, 2014*b*; Jackson, 2018; Petek and Pope, 2021). One reason for the paucity of evidence on this topic is that large-scale data linking schooling records to CJC outcomes are rare. Measuring effects can also be difficult with limited data for each individual teacher, and debate remains about the appropriateness of the standard parametric Empirical Bayes (EB) strategy used in response (Gilraine et al., 2021).

This paper seeks to make progress on both these challenges. Using nearly two decades of linked administrative schooling and criminal justice records, we develop new EB-free estimators of the variance of elementary and middle school teachers' effects on their students' future arrest, conviction, and incarceration. To study the drivers of these effects, we estimate their covariance with teachers' effects on standardized test scores, behavioral proxies for non-cognitive skills, and study skills. Doing so allows us to ask whether teachers who boost test scores, for example, also decrease their students' future CJC. A key focus is effect heterogeneity—some students may face limited arrest risk regardless of who teaches them and some teachers may excel at reaching particular students. After establishing a set of baseline, homogeneous effects, we explore the importance of accounting for heterogeneity across student characteristics and schooling environments.

The analysis is made possible by a novel merge of administrative criminal justice and education data sets in North Carolina. The combined data sets include almost two million students and 40,000 teachers. The education records cover all students in N.C. public schools in grades 3 to 12 from 1996 to 2013 and include rich data on students and their outcomes. The criminal justice data include the universe of N.C. arrests and detailed data on case outcomes, including conviction status and sentences. The data are linked by name and date of birth; comparisons of match rates to external benchmarks suggest the merge quality is high.

Our empirical strategy uses non-parametric estimators of the variance-covariance structure of teacher effects, building on the literature on variance component estimation (e.g., Krueger

and Summers, 1988; Aaronson, Barrow and Sander, 2007; Kline, Saggio and Sølvsten, 2020). The approach separates the problem of estimating the *distribution* of teacher effects from obtaining accurate estimates of any *individual* teacher's effects, side-stepping the challenges associated with the EB techniques typically employed in similar analyses and obviating the need to use "shrunken" teacher effects as explanatory variables in regressions. Our approach also easily incorporates heterogeneity along a variety of observable dimensions. The analysis relies on a rich set of characteristics available as controls to justify a conditional independence assumption, which is supported by a battery of validation tests.

Estimates of teachers' direct effects on future arrests, convictions, and incarceration are large. For example, we find a standard deviation of teacher effects on future arrests of 2.7 p.p. (11.3 percent of the sample mean) and on incarceration of 2.1 p.p. (23.6 percent). To compare these estimates to effects on outcomes studied in previous analyses, we also estimate effects on long-run academic outcomes such as high school graduation and students' plans to attend college. We find substantial teacher effects on these outcomes as well.

Teachers' impacts on short-run outcomes also vary widely. The standard deviation of teacher effects on standardized math test scores, for example, is 0.13, similar to recent estimates in other geographies (Kane and Staiger, 2008; Chetty, Friedman and Rockoff, 2014*a*). Teachers also strongly influence students' homework and reading time outside the classroom, as well as the likelihood students are suspended, come to class, and repeat a grade. While teacher impacts on test scores and study skills are tightly correlated, impacts on test scores and behaviors are not. That is, teachers who increase students' scholastic achievement are not more likely, on average, to improve students' discipline and attendance outcomes, consistent with results in Jackson (2018) and Petek and Pope (2021).

Relating short and long-run outcomes reveals that teachers who boost test scores do not meaningfully decrease their students' CJC as young adults. Shifting a student to a teacher with one standard deviation higher effect on test scores decreases their likelihood of arrest between the ages of 16 and 21 by less than 0.001 percentage points (p.p.); we cannot reject zero effect at conventional significance levels. Teachers who boost study skills have similarly limited effects on CJC. High test score effect teachers do, however, improve students' long-run academic outcomes. The estimated effect of a standard deviation shift in teacher quality on college attendance is very close to that estimated in Chetty, Friedman and Rockoff (2014*b*).

By contrast, teachers' impacts on behavioral outcomes are closely connected to their impacts on CJC. Assignment to a teacher with a standard deviation more beneficial impacts on a

summary index of discipline, attendance, and grade repetition decreases the likelihood of future CJC by two to four percent, depending on the outcome. If teacher effects on short-run behaviors are evidence of influence over non-cognitive and socio-emotional skills and traits such as conscientiousness, perseverance, and sociability (Lleras, 2008; Bertrand and Pan, 2013), the evidence supports a growing body of research suggesting that the accumulation of "soft skills" may lie at the heart of the return to education for crime (Heckman and Rubinstein, 2001; Heckman, Stixrud and Urzua, 2006; Reynolds, Temple and Ou, 2010; Heckman and Kautz, 2012; Heckman, Pinto and Savelyev, 2013; Jackson et al., 2020).[1]

While the estimates rely on non-experimental variation in teacher assignments, the detailed administrative records make it possible to control for a range of potential confounds. Multiple tests demonstrate that all effects are measured with limited bias. The estimates are insensitive to the inclusion of covariates omitted from the model, including parental education, twin indicators, and twice-lagged test scores, all of which strongly predict outcomes conditional on the baseline controls. Using teachers switches across schools and school-grades to instrument for teacher assignments, we cannot reject that estimates are unbiased. Nevertheless, we also show that our estimates provide a lower bound on the variance of causal effects and that covariances in causal effects are still identified when allowing for restricted forms of bias, including the case where estimates of test score effects are unbiased but direct effects on CJC are not.

Teachers have substantive impacts on future CJC for many types of students, including groups defined by sex, race, socio-economic status, and predicted CJC risk using all covariates. But effects are not perfectly correlated across student types. The correlation of a teacher's effects on their white and non-white students' criminal arrests is roughly 0.5, for example, indicating important heterogeneity in teachers' impacts. Effects on short-run outcomes, on the other hand, show tight correlation across groups. The correlation between a teacher's test score effects for boys and girls is 0.96, for example. The impact of assignment to a teacher good at boosting test scores or improving behaviors for observably similar students is thus similar to the average effect, since heterogeneous effects on short-run outcomes provide a poor proxy for heterogeneous effects on long-run outcomes.

Related research has shown important effects of schools themselves on students' test scores, disciplinary outcomes, and future CJC (Sorensen, Bushway and Gifford, 2019; Bacher-Hicks, Billings and Deming, 2019; Jackson et al., 2020). To examine how teachers' effects might

[1] A large literature documents the importance of non-cognitive and socio-emotional skills for long-run outcomes, including Heckman and Rubinstein (2001), Waddell (2006), Borghans, Weel and Weinberg (2008),Cunha and Heckman (2008), Cunha, Heckman and Schennach (2010), Lindqvist and Vestman (2011), Deming (2017), and Gray-Lobe, Pathak and Walters (2021).

change across different schooling environments, we extend our models to allow teachers' effects to vary by school. We also develop an estimator that only exploits *within*-school comparisons of teachers, ensuring that the estimates reflect differences in impacts on similar students and in classrooms with similar disciplinary policies, facilities, and resources. We find similar results in both exercises, with large teacher effects on CJC most tightly correlated with the impacts on behaviors rather than test scores.

To conclude, we simulate the impacts of replacing the bottom five percent of teachers based on various measures. Retention policies based on teachers' direct effects on long-run outcomes would result in large improvements, including up to 10 p.p. increases in college attendance and six p.p. reductions in criminal arrests for exposed students. Polices that target teachers using their impacts on short-run measures, however, achieve a fraction of these gains, underscoring the scope of teacher impacts not captured by these measures. Putting emphasis on different short-run outcomes trades off effects on long-run academic and CJC outcomes, with policies that emphasize test score impacts most strongly affecting the former and policies that emphasize behaviors primarily influencing the latter.

This paper contributes to a broad literature on the importance of teachers. Related research has shown that teacher quality—as measured by their influence on students' test scores—varies widely (Rivkin, Hanushek and Kain, 2005; Kane and Staiger, 2008; Chetty, Friedman and Rockoff, 2014*a*; Bacher-Hicks, Kane and Staiger, 2014*a*; Bau and Das, 2020) and has important consequences for earnings (Chetty, Friedman and Rockoff, 2014*b*). Teachers, however, impact a broad set of skills beyond those measured by standardized tests (Jackson, 2018; Jackson et al., 2020; Petek and Pope, 2021). The skills rewarded in one domain, such as the labor market, may differ from those rewarded in another, implying that what makes a teacher "good" depends on the outcome considered. Our results demonstrate the importance of teachers for a critical but understudied long-run outcome and show that the set of short-run impacts that predict these effects is strikingly different than for earnings.

We also contribute to the growing literature on whether teachers have a comparative advantage in teaching certain students (Dee, 2005; Condie, Lefgren and Sims, 2014; Gershenson et al., 2018; Delgado, 2021; Biasi, Fu and Stromme, 2021; Bates et al., 2022; Bau, 2022). We document meaningful heterogeneity in teachers' direct effects on long-run outcomes such as CJC or college attendance. However, we also find that latent teacher effects on short-run outcomes are highly correlated across students, suggesting there is limited teacher comparative advantage for students' performance on test scores, behaviors, and study skills. Identifying strong predictors of teachers' heterogeneous long-run effects is an important topic for future research.

Finally, our empirical strategy provides a new approach to studying the variance-covariance structure of latent teacher effects that does not require any EB "shrinkage" or strong parametric assumptions. Building on recent work by Kline, Saggio and Sølvsten (2020), we provide analytic standard errors that remain valid in situations where the bootstrap may be unreliable (Karoui and Purdom, 2016). Using our approach is particularly important when studying teacher effect correlations across outcomes or student types, since in general the sample variance-covariance of EB posteriors is not a consistent estimator of the variance-covariance of latent effects. While policy makers may eventually wish to obtain high-quality predictors of individual teachers' effects (e.g., for making retention decisions), obtaining accurate estimates of latent distributions is a critical input for doing so.

The rest of this paper is organized as follows. In Section 2, we describe the data and setting. In Section 3, we describe the conceptual and econometric framework used to identify and estimate teacher effects. Section 4 presents and validates the main results. In Section 5, we examine heterogeneity in teacher effects based on student characteristics and schools. In Section 6, we simulate the effects of various retention policies. Section 7 provides concluding discussion and suggested directions for future work.

2 Data and setting

In this section, we describe the administrative data used in the analysis and how the data sets are linked together. We also describe in detail how we construct our primary analysis sample and provide summary statistics.

2.1 Education records

We utilize administrative education records provided by the North Carolina Education Research Data Center (NCERDC). These data provide comprehensive records for the universe of North Carolina public school students from 1996 through 2013. Key data elements include test scores, teacher and classroom assignments, demographic characteristics of students, parents, and teachers, and disciplinary and attendance records.

2.1.1 Measuring teacher assignment to classrooms

Our analyses focus on the impacts of elementary and middle school teachers in grades four through eight. In elementary school, students are usually assigned to a single homeroom teacher, although some students have separate math and reading teachers. In middle school, students typically have separate teachers for math and reading courses. From 2006 onwards,

the NCERDC provides "course membership" files that directly link students to their teachers. Prior to 2006, we follow Rothstein (2017) and use the identity of students' end-of-year test proctor to link students to their teachers.[2]

2.1.2 Short-run outcome measures

We construct three primary measures of short-run outcomes from the NCERDC data. The first proxies for cognitive skills using scores on standardized state-wide examinations in math and reading taken by all North Carolina students. Test scores are normalized within each year and grade to have a mean of zero and a standard deviation of one in the full student population. For homeroom teachers, we use the first principal component of math and reading scores as the relevant outcome. Math and readings scores are used for math and reading teachers, respectively.

The second measure follows a large literature that uses student behaviors to proxy for non-cognitive skills (Heckman, Stixrud and Urzua, 2006; Lleras, 2008; Bertrand and Pan, 2013; Gershenson, 2016; Petek and Pope, 2021). As in Jackson (2018), we take the first principal component of standardized indicators for school discipline (primarily in- and out-of-school suspensions), days absent, and grade repetition in each year. Unlike test scores, effects on these measures may capture both changes in students' behavior and teachers' propensity to punish their students or record absences. To isolate the former component, we measure suspensions and absences the year after the student was assigned to a teacher (i.e., in $t+1$).[3] We normalize the sign of the behavioral index so that improved behaviors (e.g., fewer suspensions) corresponds to more positive values of the index.

As noted in Jackson (2018), prior research documents that these behaviors are strongly associated with important non-cognitive skills and traits (e.g., Duckworth et al., 2007). Importantly, however, teacher effects on this outcome are not intended to be interpreted as direct measures of effects on self-restraint, conscientiousness, grit, self-esteem, agreeableness, or other non-cognitive traits. Instead, the assumption is that students who improve their behaviors in school likely also improve their ability to exercise these skills. Behavioral measures therefore serve as proxies much as standardized test scores proxy for underlying fluid or crystallized intelligence. We make no attempt to correct for measurement error of these proxies for the relevant latent factors, an interesting direction for future research.

[2]Replicating this strategy in the post-2006 data confirms that proctors provide a reliable source of teacher identities.

[3]Related work also uses grades as a behavioral measure (Jackson, 2018; Petek and Pope, 2021). Since grades likely capture some of the skills also measured by test scores, we omit them from our behavioral summary measure to focus estimates on non-cognitive skills.

The third and final measure uses data on students' time spent on homework, reading, and watching television, which we interpret as proxies for students' study skills and effort. These variables are reported categorically with discretization that changes year-to-year. We convert values to hours using the mid-point of each category and normalize within grade and year. As with behaviors, we then combine the three measures into a single summary index using the first principal component.

Although test scores are available over the full panel, behaviors and study skill measures are not. Absences are available for all students beginning in 2004, and disciplinary records begin in 2001 for a subset of the schools and for all schools beginning in 2006. When estimating teacher effects on each outcome, we use all data available.

2.2 Criminal justice records

We use administrative information on arrests, charges, and sentencing from the North Carolina Administrative Office of the Courts (AOC). The data cover all cases disposed between 2006 and mid-2020 and include rich information on defendants, offenses, initial charges, convictions, and sentences. Because criminal charges in North Carolina are initially filed by law enforcement officers (as opposed to prosecutors), the charges in these data closely approximate arrests. In Charlotte-Mecklenburg County, where we have collected arrest records directly from the Sheriff, over 90 percent of arrests appear in the AOC data.[4]

The data include a large set of offenses ranging from speeding tickets to homicides. We focus our analysis on actual criminal arrests as defined by North Carolina statutes, although we also consider impacts on non-criminal traffic and municipal ordinance violations. To examine effects on the most severe categories of crimes, we also define indicators for arrest for one of the Uniform Crime Reporting program's index crimes: aggravated assault, forcible rape, murder, robbery, arson, burglary, larceny/theft, or motor vehicle theft. Throughout, we refer to outcomes in this data as indicators of "criminal justice contact" rather than crime, since arrests can occur without commission of a crime and vice versa. We focus on CJC between the ages of 16 to 21, allowing us to measure CJC for a large number of cohorts in the education data.[5]

[4]The remainder comprise non-arrest booking events recorded by the Sheriff such as federal prisoner transfers.

[5]The age of criminal responsibility in North Carolina was 16 until December 1st, 2019, when "Raise the Age" legislation increased it to 18.

2.3 Data linking

Education records were linked to criminal justice data on the basis of name and date of birth.[6] Since not all students are arrested as young adults, we do not expect 100 percent of the education records to match the criminal justice data. Comparisons of our match rates to external benchmarks suggest the link is accurate, however. Bacher-Hicks, Billings and Deming (2019), for example, estimate that 19 percent of Charlotte-Mecklenburg students are arrested between the ages of 16 and 21, a figure close to our mean rates of criminal arrest reported below.[7]

2.4 Sample construction

Following Chetty, Friedman and Rockoff (2014*a*), we treat the student-subject-year as the unit of observation. Each row in our data therefore includes a student's subject-specific outcomes (e.g., their math test scores for the math subject), their assigned teacher, their behavioral and study skill outcomes for that year, and long-run outcomes. The full data set constructed in this way includes 13 million observations. We drop teachers who appear in multiple schools (0.7 percent of records) or grades (3.7 percent) in the same year, since their students are likely only partially exposed to their potential effects, alternative and special education schools (0.1 percent), and students with an invalid contemporary or lag math and reading score, which serve as crucial controls (8.4 percent). Finally, to mitigate any potential mismatches of students to teachers, we keep teacher-subject pairs with between 15 and 100 students per year (excluding a further 6.6 percent of observations).

2.5 Summary statistics

Summary statistics for the final analysis sample are presented in Table 1. The sample includes 9,779,708 student-subject-year observations for 1,953,547 students with 39,707 different teachers. Roughly 25 percent of the sample are Black—close to the N.C. population average, 58 percent are economically disadvantaged, 22 percent have a parent with a four-year college degree, while 40 percent have a parent with a high-school education or less.

[6] We also experimented with using social security numbers, which are available for a subset of the arrest records, and found similar match rates.

[7] Other benchmarks include Cook and Kang (2016), who estimate that 6 percent of the 1987-89 N.C. birth cohorts were convicted of serious crimes between the ages of 17 and 19; Brame et al. (2014)'s analysis of the National Longitudinal Survey of Youth, who find a self-reported arrest rate between the ages of 18 and 23 of 30 percent when non-response is treated as missing at random; and data from the CJARS project (Papp and Mueller-Smith, 2021), which finds median felony conviction rates across commuting zones comparable to our CJC rates.

Test scores are normalized to be mean zero and have a standard deviation of one in the full population of students. However, in the analysis sample (Columns 1 and 2) the average math and reading test scores are slightly higher (0.061 and 0.046), primarily due to the exclusion of students without a lag score. 17 percent of the students have some disciplinary infraction in an average year and eight percent have an out-of-school suspension. The average 12th grade GPA in the sample is 3.13 (measured on a six point scale), consistent with the slight positive selection seen in test scores. About the same share of students report plans to attend a four-year college (46 percent) as students whose parents have exposure to any college (45 percent).

Contact with the justice system is prevalent. A quarter of the students have a criminal arrest between ages 16 to 21. The rate of any CJC is much higher (44 percent), with the increase driven by traffic offenses. A substantial share of the children have a serious incident of CJC between ages 16 to 21: ten percent are arrested for an index crime, ten percent are convicted of a crime, and nine percent are incarcerated (including both jail and prison). We do not observe CJC outcomes for all the students in our analysis sample. Columns 3 and 4 report statistics for the sub-sample of the students for which we do observe CJC outcome. This sample is remarkably similar to our full analysis sample.

Table 1 also reports summary statistics for the sample of children who have a criminal arrest between ages 16 and 21 (Columns 5 and 6). These students are more likely to be economically disadvantaged, their parents have less college education, and they are more likely to be male and Black. In terms of short-run measures, these children have lower academic achievement, more disciplinary infractions, and more out-of-school suspensions. They also have lower 12th grade GPA and are less likely to graduate high school at all.

3 Econometric framework

This section lays out the econometric framework we use to define and estimate teacher effects on short- and long-run outcomes and their correlation structure. We define the population parameters and estimands first, then turn to estimation. As is common in the literature, the main results use a model where teachers have homogeneous effects on all students (Chetty, Friedman and Rockoff, 2014*a*; Angrist et al., 2017), an assumption we later relax. We defer details on tests of our identifying assumptions until after we have presented the main results.

3.1 Causal and observational effects of teachers

Consider a population of students indexed by i assigned to one of J possible teachers in year t. Let Y_{ijt} denote the potential value of a generic outcome for student i if assigned to teacher j at time t.[8] Let X_{it} denote the student's observable characteristics. Potential outcomes can be decomposed into the causal effects of teachers and student observed and unobserved heterogeneity as:

$$Y_{ijt} = \underbrace{\mu_j}_{\text{Teacher effects}} + \underbrace{X_{it}'\gamma}_{\text{Observed heterogeneity}} + \underbrace{\epsilon_{ijt}}_{\text{Unobserved heterogeneity}} \tag{1}$$

where $E[\epsilon_{ijt}] = E[\epsilon_{ijt}X_{it}] = 0$ by construction. We normalize the mean of μ_j to be zero and include a constant, so that the average causal effect on the outcome of assignment to teacher j for a random student is $E[Y_{ijt}|j] - E[Y_{ijt}] = \mu_j$. Teacher effects are therefore constant over time, an assumption we relax in robustness exercises and when exploring the importance of schools. Since $E[\mu_j] = 0$, μ_j captures teacher j's effects on outcomes relative to the average teacher.

Since Equation 1 is a simple linear projection of potential outcomes onto observables, as written it imposes no additional structure on the nature of treatment effects. In the first part of our analysis, however, we follow the prior literature and assume teacher effects are homogeneous across students, allowing us to write $\epsilon_{ijt} = \epsilon_{it}$. We relax this assumption further below.

We define "observational" teacher effects as the population projection version of Equation 1 that relates actual teacher assignments to *realized* outcomes:

$$Y_{it} = \sum_j \alpha_j D_{ijt} + X_{it}'\Gamma + u_{it} \tag{2}$$

where $D_{ijt} = 1$ if student i is assigned to teacher j in year t and $= 0$ otherwise, and Y_{it} is student i's realized outcome in year t. Equation 2 is a projection defined by the population requirement that $E[u_{it}D_{ijt}] = 0 \ \forall \ j$. Observational and causal effects of teachers only coincide, however, when $E[\epsilon_{it}D_{ijt}] = 0 \ \forall j$, implying that teacher assignments are uncorrelated with unobserved determinants of potential outcomes. If this is the case, $\alpha_j = \mu_j \ \forall j, \gamma = \Gamma$, and $u_{it} = \epsilon_{it}$, and unbiased causal effects of teachers can be estimated using sample analogues of Equation 2. We call this assumption conditional independence:

[8] Our unit of observations is a student-subject-year triplet. However, to simplify the notation and exposition, we follow Chetty, Friedman and Rockoff (2014*a*) and focus on the case in which a student has only a single teacher in each year. Alternatively, i can be defined as indexing student-subject pair.

Assumption 1 *Conditional independence:* $E[\epsilon_{it} D_{ijt}] = 0 \ \forall j$.

Conditional independence does *not* necessarily require random conditional assignment of students to teachers. Instead, teacher assignments must be uncorrelated with unobserved factors that influence outcomes. This assumption rules out, for example, some teachers being systematically assigned students who are more likely to do well on standardized tests than observationally similar peers regardless of their teacher. To support this restriction, the covariates X_{it} include a large set of potential confounds. The primary estimates include year-grade-subject fixed effects; third-order polynomials in lagged math and reading test scores interacted with grade and subject; indicators for the student's academically gifted status, behavioral or educational special needs, economic disadvantage indicators, and English proficiency status; race and gender; lagged school discipline and grade repetition indicators and lagged days absent; and school and classroom means of these variables. If a variable is missing for a particular student, it is replaced with zero and a dummy variable for missing is included.

Importantly, however, Assumption 1 does not rule out some teachers being assigned students with particular observed or unobserved characteristics so long as their influence on outcomes is accounted for by the controls. This distinction is important in light of Rothstein (2010)'s influential finding that teacher assignments are correlated with observables such as twice-lagged test scores in a subset of the same North Carolina data we use. Rothstein's result is not necessarily inconsistent with Assumption 1. However, Assumption 1 does imply that adding controls for correlated observables such as twice-lagged test scores should not affect estimated teacher effects. We discuss several tests based on this idea in what follows, all of which support the validity of Assumption 1 in our application.

Despite these tests, some may naturally view conditional independence as too strong an assumption in practice given that the exact teacher assignment mechanism is unknown. We therefore also consider a weaker identifying assumption that allows for a restricted form of bias in teacher effects. Specifically, we define observational teacher effects as "forecast unbiased" if the following assumption holds.

Assumption 2 *Forecast unbiased effects:* $\mu_j = \alpha_j + \eta_j$ *and* $Cov(\alpha_j, \eta_j) = 0$.

where η_j is the difference between causal and observational effects. If observational effects were forecast unbiased and observed without measurement error, a regression of teachers' causal effects on their observational effects would yield a coefficient of one.[9] Observational

[9]In practice, α_j is estimated with error and $\hat{\alpha}_j$ is not a forecast unbiased predictor of μ_j due to measurement error even if Assumption 2 holds. We detail how we overcome this issue when testing this assumption

effects are therefore unbiased linear predictors of causal effects. Naturally, conditional independence implies forecast unbiased effects. When only the latter holds, however, any individual teacher's causal effects may be estimated with bias so long as this bias is uncorrelated with the teacher's observational effect. More concretely, Assumption 2 requires that the causal effects of teachers who appear high quality given the students they are actually assigned cannot be systematically over- or under-estimated. This restricted form of bias is plausible if high quality teachers are sometimes assigned students likely to excel no matter what, but also sometimes assigned students who face more challenges.

The discussion so far has considered a generic outcome Y_{it}. In what follows, we consider multiple short- and long-run outcomes, including math and reading test scores, proxies for non-cognitive skills such as attendance and suspensions, and future CJC. Each teacher is therefore characterized by vectors of causal and observational effects $\boldsymbol{\mu}_j = \{\mu_j^1, \mu_j^2, \ldots, \mu_j^K\}$ and $\boldsymbol{\alpha}_j = \{\alpha_j^1, \alpha_j^2, \ldots, \alpha_j^K\}$, for each of K outcomes. Likewise, Assumptions 1 and 2 can be invoked for the causal and observational effects of teachers on each outcome separately.

3.2 Parameters of interest and estimation

We focus on estimating the overall and conditional variance-covariance matrix of teachers' set of latent observational effects $\boldsymbol{\alpha}_j$. The variance of elements of $\boldsymbol{\alpha}_j$ measures how important effects are for particular outcomes. The covariance of elements $\boldsymbol{\alpha}_j$ measures how teachers' observational impacts relate across dimensions. The covariance of effects for test scores and future CJC, for example, determines whether teachers that boost scholastic achievement also reduce future criminality. Rescaling covariances by variance estimates to mimic the classic variance-covariance representation of a bivariate regression coefficient, one can easily estimate the effect of assignment to a teacher with one standard deviation larger latent effect on test scores on a long-run outcome.

Because $\boldsymbol{\alpha}_j$ are population projection coefficients, OLS estimates of $\boldsymbol{\alpha}_j$ are unbiased when the data are a random sample from the population. But they are also noisy. As a result, the variance of $\hat{\boldsymbol{\alpha}}_j$ will overstate the true variance of $\boldsymbol{\alpha}_j$. Due to correlated sampling error across outcomes, sample covariances between elements of $\hat{\boldsymbol{\alpha}}_j$ may also yield biased estimates of the covariances between elements of $\boldsymbol{\alpha}_j$.

We use variations on established approaches to obtain unbiased estimates of both latent effect variances and covariances (Kline, Saggio and Sølvsten, 2020). Our approach allows us to non-parametrically characterize the distribution of teacher effects without the use of an

below.

intermediate shrinkage step or specifying a complete statistical model. To begin, define the variance of teacher effects (for a generic outcome) as:

$$\begin{aligned} Var(\alpha_j) &= \frac{1}{J}\sum_{j=1}^{J}\alpha_j^2 - \left(\frac{1}{J}\sum_{j=1}^{J}\alpha_j\right)^2 \\ &= \left(\frac{J-1}{J}\right)\frac{1}{J}\sum_{j=1}^{J}\alpha_j^2 - 2\frac{1}{J^2}\sum_{j=1}^{J-1}\sum_{k>j}^{J}\alpha_j\alpha_k \end{aligned} \tag{3}$$

To construct our estimator of $Var(\alpha_j)$, we begin with teacher-year-level mean residuals from OLS estimates of Equation 2:

$$\begin{aligned} \bar{Y}_{jt} &= \frac{1}{n_{jt}}\sum_{i|j(i,t)=j} Y_{it} - X_{it}'\hat{\Gamma} \\ &= \alpha_j + \bar{v}_{jt} \end{aligned}$$

where n_{jt} is the number of students assigned to teacher j and time t, $\bar{v}_{jt} = \frac{1}{n_{jt}}\sum_{i|j(i,t)=j} u_{it} + X_{it}'(\Gamma - \hat{\Gamma})$, and $E[\bar{v}_{jt}] = 0$.[10] We assume $\bar{v}_{jt}$ is uncorrelated across years, which requires that any cohort or school-level shocks are independent over t.[11]

Assumption 3 *Uncorrelated teacher-year estimation error:* $E[\bar{v}_{jt}\bar{v}_{jt'}] = 0 \ \forall j, t \neq t'$

If Assumption 1 holds, then Assumption 3 can be understood as requiring that any sorting on unobservables is uncorrelated across t for each teacher. This requirement rules out sequentially correlated "runs" in unobserved student quality within teacher and is not imposed by Assumption 1 alone, which simply requires that unobservable sorting be mean zero for each teacher when averaging across all t. It is straightforward to relax Assumption 1 further, however, by assuming that it holds across t separated by at least m years: $E[\bar{v}_{jt}\bar{v}_{jt'}] = 0 \ \forall j, |t - t'| > m$. We explore sensitivity to m further below when testing for "drift" in teacher effects.

[10] $\hat{\Gamma}$ is estimated with teacher dummies as in Equation 2. This implies that the teacher-level means of $Y_{it} - X_{it}'\hat{\Gamma}$ are identical to estimates of teacher fixed effects obtained by estimating Equation 2 directly. They would not necessarily be identical if $\hat{\Gamma}$ were estimated without teacher dummies, a version of "improper" Frisch–Waugh–Lovell.

[11] Correlation in X_{it} across years for a given teacher may nevertheless violate Assumption 3 due to any estimation error in $\hat{\Gamma}$. Given the very large sample, any estimation error in Γ is likely to be small, mitigating this concern.

Under Assumption 3, an unbiased estimator of $Var(\alpha_j)$ is:

$$\widehat{Var}(\alpha_j) = \left(\frac{J-1}{J}\right)\frac{1}{J}\sum_{j=1}^{J}\binom{T_j}{2}^{-1}\sum_{t=1}^{T_j-1}\sum_{k=t+1}^{T_j}\bar{Y}_{jt}\bar{Y}_{jk} - 2\cdot\frac{1}{J^2}\cdot\sum_{j=1}^{J-1}\sum_{k>j}^{J}\bar{Y}_j\bar{Y}_k \tag{4}$$

where T_j is the number years observed for teacher j and $\bar{Y}_j = \frac{1}{T_j}\bar{Y}_{jt}$. This estimator is simply the average product of teacher-level residuals across all pairs of years. It eliminates the bias in the variance of the estimated $\hat{\alpha}_j$ by leaving out products of residuals from the same year. Similar estimators have been used in prior work to estimate the variance of teacher effects on short-run outcomes, typically by taking the average product of mean residuals across random pairs of classrooms (e.g., Chetty, Friedman and Rockoff, 2014*a*; Jackson, 2018).[12]

$\widehat{Var}(\alpha_j)$ is also numerically equivalent to the variance of the estimated $\hat{\alpha}_j$ minus a correction due to sampling variance based on the standard error of each $\hat{\alpha}_j$ (Kline, Saggio and Sølvsten, 2020):

$$\frac{1}{J}\sum_{j=1}^{J}\left[\underbrace{\left(\bar{Y}_j - \bar{Y}\right)^2}_{\substack{\text{Variance of}\\ \text{observed } \hat{\alpha}_j}} - \underbrace{\left(1-\frac{1}{J}\right)\frac{\hat{\sigma}_j^2}{T_j}}_{\substack{\text{Correction for sampling variation}\\ \text{using standard error of } \hat{\alpha}_j}}\right] \tag{5}$$

where $\bar{Y}_j = \frac{1}{T_j}\sum_{t=1}^{T_j}\bar{Y}_{jt}$, $\bar{Y} = \frac{1}{J}\sum_{j=1}^{J}\bar{Y}_j$, and $\hat{\sigma}_j^2 = \frac{1}{T_j-1}\sum_{t=1}^{T_j}(\bar{Y}_{jt} - \bar{Y}_j)^2$. Similar estimators have been used in a variety of applications, including estimates of the variance of teacher effects (e.g., Krueger and Summers, 1988; Aaronson, Barrow and Sander, 2007; Kline, Rose and Walters, 2021).

Our second object of interest is the covariance of teacher effects across outcomes. Using test score and CJC effects—α_j^A and α_j^C—as an example, this estimand is:

$$Cov(\alpha_j^A, \alpha_j^C) = \frac{1}{J}\sum_{j=1}^{J}\alpha_j^A\alpha_j^C - \left(\frac{1}{J}\sum_{j=1}^{J}\alpha_j^A\right)\left(\frac{1}{J}\sum_{j=1}^{J}\alpha_j^C\right) \tag{6}$$

Now the source of potential bias is correlated sampling error in teacher effects estimates across outcomes. Unlike variance estimation, where measurement error leads to over-dispersion, correlated measurement error across outcomes can bias covariance estimates in either direc-

[12]For example, Jackson (2018) used the estimator $E_{t,t'}\left[\frac{1}{J}\sum_{j=1}^{J}(\bar{Y}_{jt} - \bar{Y}_t)(\bar{Y}_{jt'} - \bar{Y}_{t'})\right]$ and approximated the expectation using the median value in 200 Monte Carlo simulations. Our estimator uses all possible pairs of years within a teacher instead of simulations.

tion.

Our covariance estimator is constructed assuming that Assumption 3 holds across outcomes and, much like $\widehat{Var}(\alpha_j)$, excludes products of residuals from the same year:

$$\widehat{Cov}(\alpha_j^A, \alpha_j^C) = \left(\frac{J-1}{J}\right) \frac{1}{J} \sum_{j=1}^{J} \binom{T_j}{2}^{-1} \sum_{t=1}^{T_j-1} \sum_{k=t+1}^{T_j} \bar{Y}_{jt}^A \bar{Y}_{jk}^C - 2 \cdot \frac{1}{J^2} \cdot \sum_{j=1}^{J-1} \sum_{k>j}^{J} \bar{Y}_j^A \bar{Y}_k^C \tag{7}$$

As with the variance estimator, $\widehat{Cov}(\alpha_j^A, \alpha_j^C)$ is numerically equivalent to taking the covariance of estimated effects across outcomes and subtracting a correction for within-teacher correlated measurement error:

$$\frac{1}{J} \sum_{j=1}^{J} \left[\underbrace{\left(\bar{Y}_j^A - \bar{Y}^A\right)\left(\bar{Y}_j^C - \bar{Y}^C\right)}_{\substack{\text{Observed covariance} \\ \text{across teachers}}} - \underbrace{\left(1 - \frac{1}{J}\right) \frac{\widehat{Cov}_j^{A,C}}{T_j}}_{\substack{\text{Correction for correlated} \\ \text{sampling error}}} \right] \tag{8}$$

where $\widehat{Cov}_j^{A,C} = \frac{1}{T_j - 1} \sum_{t=1}^{T_j} \left(Y_{jt}^A - \bar{Y}_j^A\right)\left(Y_{jt}^C - \bar{Y}_j^C\right)$.[13]

Comparison to alternative approaches. The prior literature often uses the covariance of EB posteriors to study how teachers' impacts along multiple dimensions or across multiple groups are related. Standard linear EB posteriors typically "shrink" observational estimates towards the overall mean with weights related to each estimate's sampling variance. The variance of EB posteriors is generally smaller than the variance of latent effects as a result of this shrinkage step, however. This can make it difficult to compare effects across groups or outcomes, since both the variance of latent effects and degree of shrinkage may be changing. In Appendix B, we show that the covariances of EB posteriors across outcomes or student groups also does not identify the covariance in latent teacher effects due to both shrinkage and correlated sampling error, and can potentially have the wrong sign. Our approach of directly estimating the variance-covariance structure of latent effects avoids these issues.

Inference. An additional benefit of our approach is that it is possible to construct analytic expressions for the sampling variances of the estimators in Equations 4 and 7, as well as their sampling co-variances. Appendix C explains how. We use this result to conduct inference instead of relying on bootstrap routines that may be misleading in high-dimensional models

[13] So far, we assumed that all outcomes are observed in all the years. In Appendix C, we present a generalization of the estimator in Equation 7 that allows for one outcome to be observed for more periods than another.

(Karoui and Purdom, 2016). Doing so also allows us to avoid the inferential complications that arise when using EB posteriors as explanatory variables in second-step regressions, as in Chetty, Friedman and Rockoff (2014*b*).

3.2.1 Interpretation under Assumption 1

When Assumption 1 holds, these estimators provide unbiased estimates of the variance covariance of causal effects of teachers across outcomes because the distribution of observational teacher effects α_j coincides with the distribution of causal teacher effects μ_j.

3.2.2 Interpretation under Assumption 2

When only Assumption 2 holds, observational variance estimates provide a lower bound on the variance of causal effects, since $Var(\mu_j) = Var(\alpha_j) + Var(\eta_j)$ (Abaluck et al., 2020). However, the difference between observational and causal *covariance* estimates—e.g., between $Cov(\mu_j^A, \mu_j^C)$ and $Cov(\alpha_j^A, \alpha_j^C)$—depends on the correlation in teacher-level bias across outcomes. For example, if biases are uncorrelated across outcomes and with underlying causal effects—e.g., $Cov(\eta_j^A, \eta_j^C) = Cov(\mu_j^A, \eta_j^C) = Cov(\eta_j^A, \mu_j^C) = 0$—the observational covariance equals the causal covariance.

Given the large set of controls for prior test scores and disciplinary infractions included in our models, it is possible that Assumption 1 holds for short-run outcomes such as test scores but not for long-run outcomes such as CJC. In this case, observational and causal covariances are related by $Cov(\alpha_j^A, \alpha_j^C) = Cov(\mu_j^A, \mu_j^C) - Cov(\mu_j^A, \eta_j^C)$. They therefore coincide whenever $Cov(\mu_j^A, \eta_j^C) = 0$, implying that sorting bias in teacher effects on future CJC is orthogonal to teachers' causal effects on test scores.[14] This expression also makes the direction of any potential bias clear. It seems plausible, for example, that the "best" teachers—i.e., with the most positive causal effects on test scores—teach in classrooms occupied by the students least likely to be arrested conditional on the controls. This pattern of sorting would make estimated observational correlations more negative than causal correlation. As we show below, however, we estimate zero correlation between observational test score and CJC effects, leaving little scope for large negative causal correlations.

4 The causal effects of teachers

This section begins by estimating teacher effects on students' future CJC. We then examine the correlation structure of teacher effects on short- and long-run outcomes. Finally, we

[14] Similar arguments are made in the case of adult earnings in Chetty, Friedman and Rockoff (2014*b*).

conduct multiple tests of Assumptions 1 and 2 to support the causal interpretation of our estimates and demonstrate the robustness of our results to alternative specifications.

4.1 Effects on CJC

We begin by estimating the impacts of teachers on students' CJC in early adulthood and comparing these estimates to effects on long-term academic outcomes. Table 2 presents variance-covariance estimates of these effects based on the empirical strategy described in Section 3. The diagonal entries reflect estimated standard deviations and the off-diagonals are correlations across outcomes. The first four columns show effects on four measures of CJC: any interaction (including traffic tickets and other non-criminal violations), criminal arrests, arrests for index crimes, and incarceration. The final three columns show effects on 12th grade GPA (measured on a 6 point scale), graduation, and plans for four-year college attendance.

Estimated teacher effects on future CJC are large. A one standard deviation increase in teacher effects would increase the likelihood of future criminal arrest, arrests for index crimes, and incarceration by 0.027, 0.018, and 0.021 p.p., respectively, or 11.25, 18.0, and 23.6 percent of the outcome mean. Teacher effects are thus larger proportionally for more severe CJC outcomes. Effects on 12th grade GPA, graduation, and college attendance are also large with, for example, an estimated standard deviation of teacher effects on the latter of roughly 0.05 p.p.

Effects on these long-run outcomes are correlated in ways one would expect. Teachers who decrease their students' odds of future CJC also make them more likely to attend college and to have better grades as seniors, consistent with the literature showing a negative relationship between years of education and CJC (Lochner and Moretti, 2004). Moreover, teachers' effects on the likelihood of future arrest are positively correlated with their effects on the probability of incarceration, as would be expected given that incarceration typically requires a preceding arrest.

4.2 Effects on short-run outcomes

Table 3 presents estimates of the variance-covariance structure of teacher effects on short-run outcomes based on the estimators in Equations 4 and 7. The diagonal entries reflect estimated standard deviations for the outcome in the row/column. The off-diagonals are estimated correlations of effects on the row/column outcomes. The top-left entry, for example, shows that the estimated standard deviation of teacher effects on test scores—combining

homeroom, math, and reading teachers—is 0.121. Since test scores are normalized to have a mean of zero and standard deviation of one in the full population of students, this means that a one standard deviation increase in teacher test score quality increases students' scores by 12.1 percent of a standard deviation on average.

The following two columns break test score effects into effects on math and reading. As in other studies, we estimate a standard deviation of teacher reading effects that is roughly half as large as teachers' math effects.[15] Because some teachers teach both math and reading either in the same or different years, we can also estimate the correlation in teacher effects on these two subjects. The estimated correlation is 0.675, implying a tight link between teacher quality in both subjects. That is, teachers who excel at increasing math test scores also tend to be high quality reading instructors, implying teaching skills generalize meaningfully across subjects.

The fourth column of Table 3 shows wide variation in teacher effects on study skills. The estimated standard deviation is 0.183. Unsurprisingly, study skills effects are correlated with effects on test scores (0.317), suggesting teachers whose students complete more homework and substitute from watching television to reading also tend to see increases in test scores.

The fifth column shows that the estimated standard deviation of teacher effects on behaviors is 0.125. Recall that the behavioral index is normalized to have a standard deviation of one in our sample, so this estimate also implies that a standard deviation increase in teacher behavioral effects improves behaviors by 12.5 percent of a standard deviation of the outcome. Interestingly, teacher effects on behaviors are only weakly correlated with teacher effects on test scores. The correlation between behavioral effects and overall test score effects, for example is 0.056. Similarly, behaviors and study skills effects are only weakly related with a correlation of 0.033.[16]

Teachers who succeed in preventing their students from acting out and skipping class, therefore, are not usually the same teachers who make students ace their standardized tests. It is possible that helping students develop skills captured by behavioral measures requires teachers to focus on different activities than those that most directly affect achievement tests. In

[15] Figure A.1 shows these estimates are comparable to results from other recent studies.

[16] Table A.1 provides a deep-dive on the specific behaviors captured by the summary index. Outcomes here are not normalized, so that, for example, teacher effects on any discipline at $t+1$ can be interpreted in percentage points. We find large variation in teacher effects on discipline and out-of-school suspensions at $t+1$, with a standard deviation of effects on the latter falling at 0.026. Teachers meaningfully affect grade repetition and $t+1$ absences as well, with effect standard deviations of 0.008 and 0.759, respectively. Despite differences in normalization and definitions, these estimates appear roughly comparable to those in the recent literature.

classrooms where students are at risk of suspension or skipping school, teachers may opt to focus on the former at the expense of the latter. While perhaps surprising, similar results have been found in other contexts. Jackson (2018) and Petek and Pope (2021) find a correlation of 0.15 between teacher value-added on a behavioral index and test scores.[17]

4.3 Connecting short- and long-run effects

How are short- and long-run effects related? One simple summary statistic is the coefficient from a population regression of teacher effects for long-run outcome C on short-run effects for outcome A, or $\frac{Cov(\alpha_j^C, \alpha_j^A)}{Var(\alpha_j^A)}$. Using the estimators in Equations 4 and 7, it is straightforward to obtain a plug-in estimate of this object. Figure 1 reports these estimates for the long- and short-run outcomes studied above. Each coefficient has been re-scaled by the standard deviation of short-run outcome effects, so that they can be interpreted as the implied impact on the long-run outcome of exposing a student to a teacher with one standard deviation higher effects on the short-run outcome.[18]

The signs of effects are normalized so that the hypotehtical change always results in an improvement in the short-run outcome (e.g., higher test scores, fewer suspensions). The bars are grouped by short-run outcome, with each bar showing the estimated effect on each long-run outcome. The figures above the bars report effects as a percentage of the outcome's baseline mean. Because we measure the variance-covariance structure of *latent* teacher effects, these estimates reflect impacts of assignment to teachers whose *actual* impacts on the short-run outcome are one standard deviation higher. We return to the costs of needing to estimate individual teachers' effects in finite samples in the last section of the paper.

Panel (a) presents the results for any CJC, criminal arrests, arrests of index crimes, and incarceration. Consistent with the small estimated correlations, shifting students to teachers better at increasing test scores has limited impacts on future arrests. Effects on any CJC and criminal arrests are small, with confidence intervals that include zero. Effects on arrests for index crimes and incarceration are larger but still no more than 0.7 percentage points. Teachers' effects on study skills show similar patterns, with effects statistically indistinguish-

[17] Petek and Pope (2021) is the only other estimate, to our knowledge, of the correlation between teacher effects on study skills and test scores or a behavioral index. Interestingly, they find the opposite pattern: a strong correlation between teacher effects on learning skills and behaviors (0.459) and a weaker correlation between learning skills and test score teacher value-added (0.174). Study skills are measured in different ways in Petek and Pope (2021), possibly explaining the differences.

[18] Table A.2 reports estimates of the correlations between effects on pairs of short- and long-run outcomes underlying these coefficients.

able from zero. For the purposes of recruiting, retaining, or rewarding teachers likely to help their students avoid criminal careers, therefore, test score and study skill value-added is not likely to be particularly useful metric.

Unlike CJC outcome, we find that teachers who increase test scores do increase students' 12th grade GPA, their likelihood of graduation, and their plans to attend four-year college. The estimated effect of a one standard deviation shift in teacher quality on the latter is roughly 1.3 percent (roughly 0.6 p.p., similar to the estimated impact on actual college attendance in Chetty, Friedman and Rockoff (2014*b*) of 0.86 p.p.). Table A.3 shows that we find similar patterns when using conventional methods that regress students' future CJC on assigned teachers' value-added estimated in a multi-step EB procedure. The effect of assignment to a teacher with one standard deviation higher test score value-added on future criminal arrests, for example, is 0.08 p.p.

By contrast, Panel (b) shows that exposing students to teachers with more positive effects on behaviors has a large impact on future CJC. A one standard deviation shift in behavioral effects is associated with a 2.3 percent decrease in the likelihood of a criminal arrest and a three percent reduction in the likelihood of being incarcerated between ages 16 to 21. Teacher quality measured through their impacts on these outcomes, therefore, is very relevant for improving students' long-run criminal justice outcomes.[19] As with test scores, Table A.4 shows results change little when using alternative regression-based estimators. Teachers' impacts on behaviors also affect long-run academic outcomes, with similar effects as test score impacts on 12th grade GPA, for example.[20]

4.4 Multivariate relationships between teacher effects

Estimates of the variance-covariance structure of teacher effects can be used to estimate the infeasible regression of teachers' effects on CJC on all of their short-run effects simultane-

[19]Part of the long-run effects of exposure to teachers with positive effects on behaviors may flow through development of certain skills and part may flow through the impacts of the behavior itself. Bacher-Hicks, Billings and Deming (2019), for example, find that assignment to schools with more strict discipline policies results in more criminal justice contact. Sorensen, Bushway and Gifford (2019) report similar findings. Consistent with our results, Bacher-Hicks, Billings and Deming (2019) also find that schools that improve test scores do not impact future arrests or incarceration.

[20]Figure A.2 explores which specific behaviors drive the results using the summary index. For both all and criminal CJC, effects on school discipline matter most. Attendance effects are also meaningfully correlated with future CJC. Grade repetition is largely orthogonal, perhaps suggesting that this outcome is more closely connected to academic achievement measures such as test scores.

ously:

$$\alpha_j^C = \beta_0 \alpha_j^A + \beta_1 \alpha_j^B + \beta_2 \alpha_j^S + e_j$$

where superscripts A, B, and S indicate effects on test scores, behaviors and study skills, respectively. $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \beta_2)'$ is straightforward to calculate given variance-covariance estimates, since:

$$\boldsymbol{\beta} = E[(\alpha_j^A \ \alpha_j^B \ \alpha_j^S)' \cdot (\alpha_j^A \ \alpha_j^B \ \alpha_j^S)]^{-1} E[(\alpha_j^A \ \alpha_j^B \ \alpha_j^S)' \alpha_j^C]$$

Table 4 presents estimates of $\boldsymbol{\beta}$. Consistent with Figure 1, "horse-racing" the short-run effects shows that the key predictor of teachers who reduce CJC is teachers' effects on behaviors. Test score effects have negligible impacts on future CJC, while behavioral effects have much larger ones. For 12th grade GPA, graduation, and college attendance, both effects matter independently and enter with substantial regression coefficients.

Given estimates of the total variation in effects on long-run outcomes from Table 2, it is straightforward to calculate the implied R^2 from these regressions. For criminal arrests, the R^2 is 0.042, while for college attendance it is 0.02. Thus only a small share of the total variance in teacher effects on long-run outcomes is jointly explained by all short-run effects. This result implies that while behavioral effects are strongly correlated with criminal arrests, teachers also impact CJC in many ways orthogonal to their impacts on suspensions, attendance and grade repetition. The same is true to an even greater degree for 12th grade GPA, high school graduation, and college attendance. Any policy focused on these short-run outcomes will therefore likely neglect substantial heterogeneity in teachers' importance for each of these long-run outcomes.

4.5 Validating effects

In this section, we present multiple additional analyses that support the casual interpretation of our previous estimates for both short- and long-run outcomes. The robustness analyses include tests for omitted variable bias, checks for forecast unbiasedness, analyses that relax Assumption 1 and allow for unrestricted school effects, and investigations of the sensitivity of results to different specification and modeling choices.

4.5.1 Omitted variables tests of Assumption 1

The causal effects of teachers defined in Equation 1 are invariant to the inclusion of additional controls in the model. A natural test of Assumption 1 therefore assesses the sensitivity of observational effects of teachers to the inclusion of controls excluded from the original model, denoted W_{it}. Specifically, consider the augmented "long" regression model given by:

$$Y_{it} = \sum_j \tilde{\alpha}_j D_{ijt} + X'_{it}\tilde{\Gamma} + W'_{it}\rho + \tilde{u}_{it} \tag{9}$$

The canonical omitted variable bias formula implies that the sensitivity of $\hat{\alpha}_j$ to the omission of W_{it} is identified by a regression of $W'_{it}\rho$ on D_{ijt}. Likewise, the sensitivity of the relationship between $\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$ and outcomes is identified by a regression of $W'_{it}\rho$ on $\hat{\alpha}_{it}$.[21] Critically, it must be that $Var(W'_{it}\rho) > 0$, otherwise such tests will show no sensitivity mechanically. We show below, however, that our omitted variables are strongly predictive of outcomes conditional on the regular controls X_{it}.

Results. Figure 2 depicts the correlation between estimated teacher effects ($\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$) and predicted outcomes ($\hat{Y}_{it} = W'_{it}\hat{\rho}$) using twice-lagged test scores, parental education, and family fixed effects for twins as omitted variables.[22] Estimates of teacher effects come from OLS estimates of Equation 2 with our standard set of controls. Estimates of ρ come from OLS estimates of Equation 9.

For all outcomes, teacher effects are uncorrelated with predicted outcomes. The magnitude of each slope coefficient is extremely small. The slope coefficient for any criminal arrest between ages 16-21, for example, is 0.00135. This estimate implies that the impact of a one standard deviation increase in teacher effects on future arrests ($\sigma^y = 0.027$—see Table 2) may be biased by $0.027 \cdot 0.00135 = 0.000036$ due to omitted variables. Similar results hold for test scores, behaviors, and academic long-run outcomes (see Figure A.3 and Tables A.5, A.6, and A.7). Though these tests include all students, results change little when regressing $\hat{Y}_{it}$ on $\hat{\alpha}_{it}$ in the sample of twins only.

Columns 1, 3, and 5 of Table A.5 demonstrate that these omitted variables strongly predict

[21] These are the "short" regressions. Any sensitivity of observational estimates to omitted variables may occur due to either sorting bias or heterogeneous effects of teachers. Including additional controls in the model implicitly changes the conditional variance of D_{ijt} and the types of students—in terms of their X_{it}—given most weight in estimating each teacher's effects. Although we find that our primary estimates are not sensitive to omitted variables, we extend the model to allow for potential heterogeneity in teacher effects in the final part of the paper.

[22] Non-twins are also included and grouped into a single fixed effect.

test scores, behavioral measures, and study skills.[23] Table A.6 reports analogous estimates for long-term CJC outcomes. The patterns are similar. Reassuringly, the omitted variables are especially predictive of criminal arrests, our main outcome of interest. Including them in the teacher effect specification increases the R^2 from 0.089 to 0.107, a 20 percent increase in the model's explanatory power.

We emphasize that the identifying assumption in our model is not that teachers are conditionally randomly assigned. Instead, Assumption 1 requires only that teacher assignments are conditionally mean independent of the relevant unobservables. Although Rothstein (2010) shows that teacher assignments are correlated with twice-lagged scores, the preceding exercises show that including these variables in the model does not impact estimated teacher effects, consistent with Assumption 1 and the arguments in Chetty, Friedman and Rockoff (2016, 2017) and Jackson (2018).

4.5.2 Instrumental variable tests of Assumptions 1 and 2

Define the population projection of teachers' causal effects onto observational effects as:

$$\mu_j = \lambda \alpha_j + \eta_j$$

Assumption 1 implies that $\lambda = 1$ and $\eta_j = 0 \; \forall j$. Assumption 2 implies only that $\lambda = 1$. With an appropriate instrument, it is possible to test whether $\lambda = 1$, a sufficient condition for Assumption 2 and a necessary condition for Assumption 1. To see how, consider the relationship between estimated observational effects and outcomes implied by the causal model:

$$Y_{it} = \lambda \hat{\alpha}_{it} + X'_{it}\gamma + \epsilon_{it} + \eta_{it} + \lambda \xi_{it} \tag{10}$$

where $\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}, \hat{\alpha}_j = \alpha_j - \xi_j, \xi_{it} = \sum_j \xi_j D_{ijt}$, and $\eta_{it} = \sum_j \eta_j D_{ijt}$.

OLS estimates of λ are inappropriate because $\hat{\alpha}_{it}$ may be correlated with ϵ_{it}, η_{it}, and ξ_{it}. In fact, if $\hat{\alpha}_{it}$ is estimated in the same data as Equation 10, $\hat{\lambda} = 1$ mechanically. However, given an instrument Z_{it} that is relevant (i.e., $Cov(Z_{it}, \hat{\alpha}_{it}) \neq 0$) and excludable (i.e., $Cov(Z_{it}, \epsilon_{it}) = Cov(Z_{it}, \eta_{it}) = Cov(Z_{it}, \xi_{it}) = 0$), it is possible to estimate λ using 2SLS.[24]

[23] Columns 2, 4, and 6 report the regression coefficients underlying Figure 2.

[24] Angrist et al. (2017) develop related tests for bias in observational estimates of school effects using lottery-based admissions offers. Since we use a single instrument, our test is equivalent to the "omnibus" test for bias they propose. Abaluck et al. (2020) exploit plan termination to test for forecast bias in observational estimates of mortality differences across health insurance plans. As noted above, $\lambda = 1$ both when effects are unbiased and when they are only forecast unbiased.

We use teacher switches across schools and grades to develop instruments. Intuitively, these test ask whether teachers' observed impacts on outcomes when they enter a new school or school-grade match what we would predict based on their impacts in other data. To define the instrument, let E_{it} be an indicator for whether a new teacher enters school-grade $sg(i,t)$. Let $\tilde{\alpha}_{sgt}$ be the mean of $\hat{\alpha}_j$ for all new teachers in sg and time t, where $\hat{\alpha}_j$ is estimated using all school-grades except sg. The instrument at the school-grade level is $Z_{it} = E_{it}\tilde{\alpha}_{sg(i,t)t}$ and is defined analogously at the school level.[25]

We assume that new teacher entry is uncorrelated with student unobservables, or $Cov(Z_{it}, \epsilon_{it})$. Because the instrument is defined at the school-grade (or school) level, any within school-grade (or school) sorting is not a concern. This assumption rules out, however, teachers with higher estimated effects systematically entering schools or school-grades where students are more likely to excel on average.[26] We also assume that the instrument is uncorrelated with teacher-level bias (i.e., $Cov(Z_{it}, \eta_{it}) = 0$) and estimation error in teacher effects (i.e., $Cov(Z_{it}, \xi_{it}) = 0$). Estimating $\tilde{\alpha}_{sgt}$ using all school-grades beside sg bolsters this assumption.

Results. Table 5 reports estimates of λ using teacher switches at the school and school-grade level. Panel (a) shows that for all short-run outcomes we cannot reject $\lambda = 1$. For test scores, for example, the point estimate using teacher switches across school-grades is 1.002 (0.012). Estimates for behavioral measures and study skills are similar, but slightly less precise due to the shorter panel over which they are observed. Estimates for teachers' direct effects on long-run outcomes in Panel (b) are likewise consistent with no bias. In each case, we cannot reject $\lambda = 1$, although λ is less precisely estimated than for short-run outcomes.[27]

The appendix contains several variations on Table 5 that probe the robustness of these results. Table A.8, for example, demonstrates that we also cannot reject unbiased effects when school-grade fixed effects are included, so that only variation in which teachers are assigned to a given school-grade is exploited. Table A.9 shows the sensitivity of our estimates of λ for our primary long-run outcome, any criminal arrests, with increasingly fine-grained

[25] Chetty, Friedman and Rockoff (2014*a*) and Bacher-Hicks, Kane and Staiger (2014*b*) exploit changes in estimated teacher effects and changes in outcomes within a school-grade to estimate λ. This approach is equivalent to stacking the data for each pair of consecutive years, controlling for school-grade-pair effects, and using the interaction of school-grade indicators and indicator for the second year in each pair as the instrument. Since this approach exploits many instruments, an important concern is whether a weak first stage may bias estimates towards the OLS estimate of 1.

[26] This assumption need hold only conditional our standard student-level controls, as well as additional ones such as district-grade-year fixed effects.

[27] The large first-stage F-statistics reported at the bottom of the table also indicate that the instruments induce substantial variation in exposure to high and low quality teachers.

sets of fixed effects, and demonstrates that the instrument is uncorrelated with predicted outcomes based on parental education and twice-lagged test scores.[28]

4.5.3 Are teacher effects actually school effects?

To show that our estimates indeed reflect the causal effects of teachers and are not confounded by omitted school effects, we conduct two complementary analyses. The first allows for arbitrary sorting of students to teachers within a school but assumes that any selection in teacher assignment within a school is uncorrelated across schools. The second allows for arbitrary sorting of teachers across schools but assumes that assignment of teachers to students within a school is conditionally random (Assumption 1). Both analyses yield results similar to our main estimates, supporting the causal interpretation of our teacher effect estimates.

Our first robustness analysis estimates the variance-covariance of teacher effects using teachers who switch schools and exploits the relationship between their short- and long-run effects *across* schools. If school effects drove our estimates, we would expect meaningful attenuation, since this approach effectively asks whether teachers who improve behaviors in school A reduce CJC in school B, or whether teachers who improve test scores in school A also do so in school B etc.

The variance-covariance estimators are analogous to those in Equations 4 and 7:

$$\left(\frac{J-1}{J}\right)\frac{1}{J}\sum_{j=1}^{J}\binom{S_j}{2}^{-1}\sum_{s=1}^{S_j-1}\sum_{k=s+1}^{S_j}\bar{Y}_{js}\bar{Y}_{jk} - 2\cdot\frac{1}{J^2}\cdot\sum_{j=1}^{J-1}\sum_{k>j}^{J}\bar{Y}_j\bar{Y}_k \tag{11}$$

where S_j is the number of schools that teacher j teaches at during the sample period and $\bar{Y}_{js}$ is the teacher's mean residual in school s, or $\frac{1}{n_{js}}\sum_{t|s(j,t)=s}\sum_{i|j(i,t)=j} Y_{it} - X'_{it}\hat{\Gamma}$. Only teachers who move across school, i.e., those with $S_j \geq 2$, are included.

In addition to testing for omitted school effects, the estimator defined in Equation 11 significantly weakens our identifying assumptions by allowing for arbitrary sorting of students to teachers within a school. It rules out however, common sorting across schools, such as a scenario in which students who are more likely to excel on standardized tests are assigned to teacher j both in school A and in school B.

Results. Table A.10 reports estimates of infeasible regressions of long-run latent teacher

[28] Although Rothstein (2017) argues that teacher switches in North Carolina are correlated with student preparedness, our tests only require switches to be conditionally orthogonal to unobserved determinants of outcomes. Table A.9 shows that this holds for CJC.

effects on short-run effects based on this approach. As in the primary estimates, test score effects are weakly related to future CJC, unlike behavioral effects. The impact of a one standard deviation increase in teacher behavioral quality is similar to that of estimates that utilize all variation.

The overall standard deviations of teachers' direct effects on long-run outcomes are slightly smaller for some outcomes, but still large. Effects on any criminal arrest, for example, have an estimated 1.9 p.p. standard deviation relative to 2.7 in estimates leaving out a year rather than a school. As in the primary estimates, however, all short-run effects continue to explain a relatively small share of the variance in long-run effects (<11 percent).[29] Moreover, among the set of teacher who switch schools, Figure A.4 shows that variance and covariance estimates of teacher effects using the primary estimator and this between-school version have a correlation of 0.99. Omitted school effects are therefore unlikely to explain our conclusions.

Our second robustness analysis estimates the variance-covariance of latent teacher effects on all outcomes separately for each school using Equation 4 and computes the weighted average. These estimates exploit purely within-school variation, only comparing the impacts of teachers working in the same environments. Any potential school effects would thus be washed out.

Results. Table A.11 reports the implied regression coefficients summarizing the relationship between short- and long-run effects using this approach. Estimates are very similar to those using across-school variation in teacher effects in Table 4. Teachers' effects on behaviors strongly predict long-run CJC outcomes such as criminal arrest and incarceration. Moreover, teachers' effects on test scores are much less predictive of future CJC, with coefficients ten times smaller than those on behavior. The total variance of teachers' direct effects on each long-run outcome is naturally lower, reflecting the fact that we have excluded all between-school variation.

4.5.4 Specification robustness

To explore how sensitive our results are to modeling choices, we estimate a large number of specifications using 811 different potential sets of controls and, in each case, construct estimates of the impact of a one standard deviation increase in a teacher's test score and behavioral effects on the likelihood of a future criminal arrest.[30] All models include lag

[29] Because the set of teachers who switch schools may be different then the overall population, there is no reason to expect direct effect variances to be identical to the primary estimates.

[30] Study skills effects are omitted for brevity and to speed computation.

third-degree polynomials in math and reading scores interacted with grade, as well as year-grade-subject fixed effects. Other possible controls include school, school-grade-year, or school-grade-classroom-year means of other included covariates, lag absences and discipline, exceptionality and gifted indicators, limited English proficiency status, gender and race, parental education, grade repetition, and twice-lagged test scores. The results reported in Figure A.5 show that our preferred specification is not an outlier. For test scores effects, our preferred estimate is close the median estimate found when including most potential controls. For behavioral effects, the estimate is among the most conservative possible.

5 Heterogeneous effects

The preceding analysis assumes that teacher effects are the same across students and schools. It is possible, however, that some teachers excel at reaching particular types of students or adjust their teaching priorities based on the classroom environment. Teachers' impacts on particular types of students and in particular schools may thus differ from their average effects. To examine this question, we extend the model in Equation 2 to allow for heterogeneous teacher effects:

$$Y_{ijt} = \sum_j \left(\mu_j + U_{it}'\beta_j\right) D_{ijt} + X_{it}'\Gamma + \epsilon_{it}$$

where U_{it} is a subset of X_{it}, such as race, gender, or socio-economic status normalized to be mean zero. Teacher effects depend on these observables, with $\mu_j(u) = \mu_j + u'\beta_j$ denoting teacher j's effects on students with observables u. Estimating this model under Assumption 1 allows us to estimate the variance-covariance structure of teachers' effects within and across groups.

5.1 Heterogeneity based on student characteristics

We focus on four sets of student characteristics: white vs. non-white students, boys vs. girls, students who are economically disadvantaged vs. not, and student with above vs. below median predicted arrest risk. Table A.12 presents averages of student characteristics and short- and long-run outcomes for these groups. A few disparities are worth noting. White students are meaningfully less likely to have CJC than non-white students, including a seven p.p. (33 percent) lower likelihood of a criminal arrest and 3.7 p.p. (51 percent) lower likelihood of being incarcerated. Similarly, girls and students from higher socioeconomic backgrounds have lower CJC rates.

Wide variation in the incidence of CJC suggests that teacher effects on CJC may vary considerably across groups. Table 6 presents estimates of teacher effects on CJC outcomes for different students. Effect variances are often larger for groups with a higher prevalence of CJC, but remain substantial for all student types and for both moderate and more serious contact types. Examining columns 1 and 2, for example, shows that teacher effects are more dispersed for non-white than for white students. Estimates are similar, however, for some groups with large differences in baseline CJC rates. The standard deviation of teacher effects on criminal arrests is 2.72 p.p. for economically disadvantaged students and 2.97 p.p for non-disadvantaged students, despite a nearly 100% difference in average arrest rates.

Even if teacher effects vary widely in multiple sub-populations, any individual teacher's effects may differ across students. Figure 3 examines this possibility by plotting the estimated correlation of teachers' effects across student types. For test scores, we find surprisingly high correlations for all groups. Teachers' test score effects on boys vs. girls, white vs. non-white students, students who are economically disadvantaged vs. not, and student with above vs. below median predicted arrest risk all have correlations about 0.9. Teachers' effects on cognitive outcomes therefore largely generalize across a wide variety of students. Teachers' effects on study skills show a similar pattern. Effects on behaviors are also strongly correlated, although less so. For example, the correlation of effects for boys vs. girls is roughly 0.75. Hence teacher quality for promoting non-cognitive skills also generalizes to a large degree across groups. Good teachers, as measured by short-run outcomes, thus appear to largely be good teachers for everyone.

Panels (b) and (c) of Figure 3 show that teachers' direct effects on long-run outcomes, however, display much weaker correlations for some groups. The correlation of teacher effects on white and non-white students' criminal arrests, for example, is roughly 0.5. As noted earlier, teacher effects on *all* short-run outcomes explain a small share of the variation in effects on long-run outcomes. Teachers' impacts on students through channels potentially uncorrelated to short-run outcomes are therefore highly heterogeneous. This finding is thus also consistent with studies that estimate meaningful effects of matching students to same-race teachers (e.g., Gershenson et al., 2018), who may be better at promoting skills unreflected in these short-run outcomes.

Finally, we connect these estimates by calculating the implied effects on long-run outcomes of exposing students to teachers with higher student type-specific quality. Since short-run effects are so correlated across groups, increasing student-specific quality largely implies exposing students in both groups to the same teachers. In particular, tight correlations in teacher test score effects across groups implies the impacts of heterogeneous test score quality

on future CJC is similar to the null average effect documented earlier, as shown in Figure A.6.

Still, Figure 4 shows that there is some evidence of heterogeneity for the impacts of quality measured by student-specific effects on behaviors. Panel (a) shows larger effects effects on criminal arrests for white students and economically disadvantaged children, though confidence intervals overlap. Effects on boys and girls are similar, but when compared to outcome means the impacts on girls is twice as large. Panel (b) reports effects on future incarceration, where there is less heterogeneity in effects. Despite some differences, in almost all cases for both outcomes we cannot reject that these potential heterogeneous impacts are in fact the same across student types.

5.2 Heterogeneity by schooling environment

A final dimension of heterogeneity we explore is schools, which have important impacts on students' outcomes through peers, staff resources, facilities, disciplinary policies, and other channels (Cullen, Jacob and Levitt, 2006; Billings, Deming and Rockoff, 2013; Beuermann et al., 2018; Bacher-Hicks, Billings and Deming, 2019; Billings, Deming and Ross, 2019; Sorensen, Bushway and Gifford, 2019; Abdulkadiroğlu et al., 2020; Jackson et al., 2020; Bruhn, 2020; Bau, 2022). It is possible that teachers adjust the skills they focus on in the classroom in response to the school environment. In some settings, teachers may be able to simultaneously focus on skills that both increase test scores and decrease CJC, for example.

To explore this question, we treat the identity of the school the student attends as an additional observable dimension of effect heterogeneity by estimating:

$$Y_{ijt} = \underbrace{\sum_{j} \left(\alpha_j + H'_{s(i,t)}\beta_j\right) D_{ijt}}_{\text{School specific teacher effects } (\alpha_{js})} + X'_{it}\Gamma + \epsilon_{it} \tag{12}$$

where $H_{s(i,t)}$ is a vector of length S (the number of schools in the data) that indicates the school student i attends in year t.

Results from this model are similar to overall estimates. To focus on estimates that hold the schooling environment constant, we use the within-school estimator of covariances introduced in Section 4.5.3. Table A.13 reports estimates of the implied regression coefficients. The results are similar to those in Table 4, the primary estimates, and Table A.11, which exploits within school comparisons but assumes homogeneous teacher effects. Among short-

run measures of teacher quality, effects on behaviors remain the strongest predictor of future CJC, while effects on test scores have a negligible impact.

6 Implications for teacher retention policies

There has been substantial discussion of how estimates of teachers' impacts on student outcomes can be incorporated into teacher retention decisions (Rothstein, 2010; Hanushek, 2011; Neal, 2011; Chetty, Friedman and Rockoff, 2014*a*). Value-added-based evaluations have already been implemented in certain jurisdictions (Biasi, 2021). Ideally, a district would evaluate teachers based on their impacts on students' long-run well being. Doing so is not typically feasible, however, since long-run outcomes are by definition not observed for many years. As a result, in practice teachers are evaluated using their impacts on short-run outcomes, such as test scores. Moreover, since teacher's *true* impacts on short-run outcomes are not observed, districts must use estimated effects to make decisions.

To demonstrate the implications of our findings for policy, we compare the potential impacts of policies that replace the worst-performing teachers based on various measures with an average teacher. Since there are multiple relevant long-run outcomes, we construct possibility frontiers that trade off potential gains on long-run academic and CJC outcomes by placing different emphasis on different measures of teacher quality. To provide a simple connection between our variance estimates and the relevant quantities for these simulations, throughout this section we assume that teacher effects are normally distributed.[31] Such parametric assumptions are not necessary (Gilraine et al., 2021), but relaxing them is also unlikely to affect the main conclusions from our analysis in this section.

We begin with the ideal (and infeasible) measures that directly capture teachers' effects on long-run outcomes. Specifically, consider a district that seeks to increase college attendance and reduce criminal arrests. As demonstrated above, teachers who increase the former are not necessarily those that reduce the latter. Denote teacher effects on college attendance by μ_j^A and on future criminal arrests by μ_j^C. The long-run score is a simple weighted average:

$$\text{Index}_j^{\text{long-run}} = \omega\mu_j^C + (1-\omega)\mu_j^A \tag{13}$$

where $\omega \in [0, 1]$.

By varying ω, it is straightforward to trace out potential gains in each outcome from replacing the five percent of teachers with the worst score. The rightmost dotted curve in Figure 5

[31] Appendix D presents more details on the calculations of each policy counterfactual.

reports the results of this exercise. If long-run effects were directly observed, the district could achieve increases in college attendance of up to 10 p.p. and decreases in criminal arrests of up to 5 p.p. for exposed students. Naturally, increasing one outcome requires reducing effects on the other. Where a district should locate on this frontier depends on their preferences over long-run outcomes.

Since teachers' effects on long-run outcomes are not observed, these estimates represent an upper bound for improvements that any policy can achieve. In practice, districts rely on teacher's impacts on short-run outcomes to proxy for teacher quality. The next set of lines in Figure 5 demonstrates feasible gains from doing so. We construct a weighted index of teacher effects on study skills, behaviors, and test scores:

$$\text{Index}_j^{\text{short-run}} = \omega_1 \mu_j^T + \omega_2 \mu_j^B + (1 - \omega_1 - \omega_2)\mu_j^S \tag{14}$$

and examine the impacts on long-run outcomes of replacing the bottom five percent of teacher with the average teacher according to $\text{Index}_j^{\text{short-run}}$ for different values of $\omega_1 \in [0, 1]$ and $\omega_2 \in [0, 1]$, where $\omega_1 + \omega_2 + \omega_3 = 1$.

The red dashed line in Figure 5 reports the results of these exercises. Using effects on short-run outcomes, the district could achieve increases of nearly 2 p.p. in college attendance and decreases of no more than 2 p.p. in criminal arrests for exposed students. Thus, while there are still meaningful potential improvements in long-run outcomes, the frontier lies far to the interior of the infeasible policy.

The green (triangle), blue (square), and purple (circle) points show the effect of placing full weight on study skills, behaviors, or test scores, respectively. The blue square shows that scores that maximize impacts on future CJC place almost full weight on behavioral outcomes. The green triangle shows that scores that maximize impacts on college attendance place significantly more emphasis on test scores. Though close to the frontier, the triangle is slightly to the interior, demonstrating that even if the district sought to increase college attendance as much as possible, they would place at least some weight on behaviors.

In practice, even teacher effects on short-run outcomes are not directly observed and must be estimated instead. The red dashed-line therefore reflects what could be achieved with the best possible estimates, i.e., that coincide with the truth. The costs of estimating scores instead will depend on the information the district has available—how many years teachers are observed for, for how many students they teach, etc.—and the quality of the models they use to predict teacher effects on short-run outcomes.

To illustrate the potential loses from estimating instead of observing teacher effects on short-run outcomes, we adopt common Empirical Bayes methods proposed in the value-added literature (e.g., Kane and Staiger, 2008; Chetty, Friedman and Rockoff, 2014*a*; Gilraine et al., 2021). These results are shown in the solid orange line in Figure 5. Naturally, only a portion of the gains from using true effects on short-run outcomes are achievable when these effects must be estimated. Using our data, the cost implies reductions in arrests or improvements in college attendance that are roughly half as large.

7 Conclusion

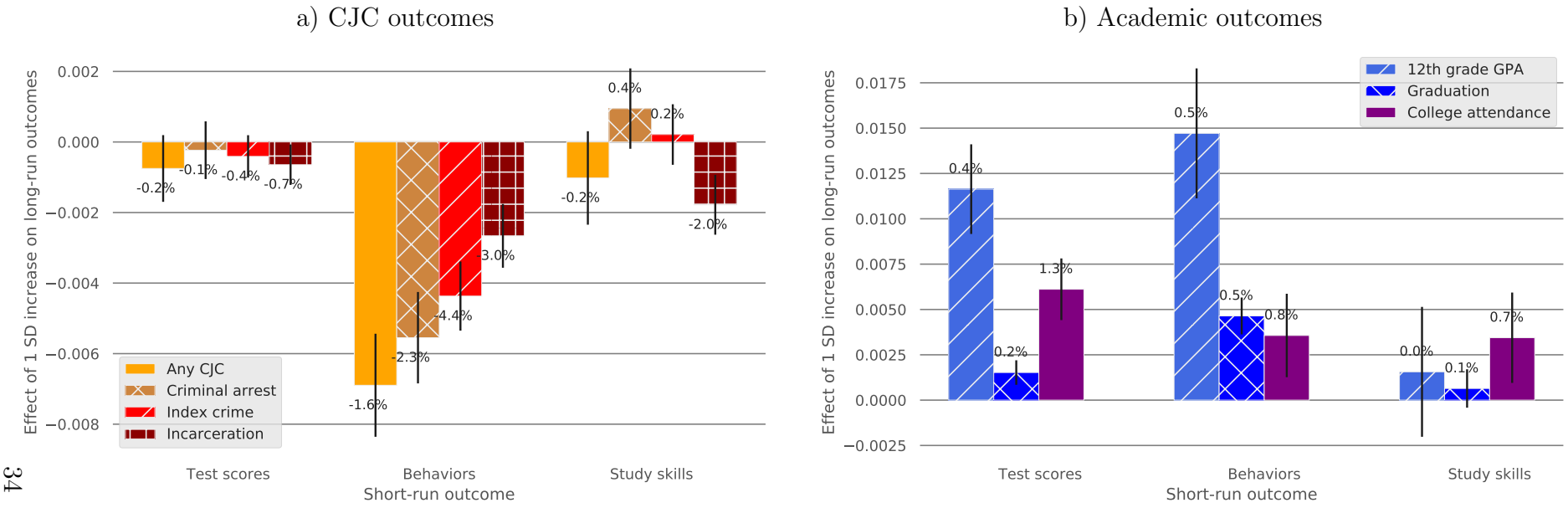
Teachers help students develop a variety of skills necessary to be successful, healthy, and happy adults. The skills needed to excel in one aspect of life, such as the labor market, may differ from those needed in another, such as avoiding entanglement in the criminal justice system. Although prior work demonstrates that teachers who increase students' cognitive skills captured by standardized tests scores increase their college attendance and adult earnings, we find that teachers' test score impacts are orthogonal to students' criminal justice contact as young adults. One of the most common and wide-spread measures of teacher quality is thus irrelevant for an outcome with life-changing consequences for large share of the population (Brame et al., 2012).

Instead, teachers who improve proxies for non-cognitive skills such as rates of school discipline and attendance have meaningful impacts on students' future arrest, conviction, and incarceration rates. These same teachers also increase students' long-run academic outcomes, including college attendance plans and 12th grade GPA. These results underscore that development of these non-cognitive skills is crucial for a wide range of outcomes, but especially CJC. They are consistent with a growing number of studies showing that educational policies and interventions that decrease CJC often primarily operate through development of these non-cognitive channels (Deming, 2009, 2011; Heckman, Pinto and Savelyev, 2013; Heller et al., 2016).

Taken together, however, teacher effects on all short-run outcomes explain a small share of their direct effects on CJC, which also exhibit substantial heterogeneity across groups. In other words, teachers' impacts on their students lives are complex and not often captured by their impacts on short-run outcomes. Though effects on test scores and behaviors can serve as useful measures of teacher quality, future research should continue to develop tools to measure teacher quality dimensions orthogonal to these outcomes. Retention and incentive policies based on richer models of teacher quality are likely to be substantially more powerful

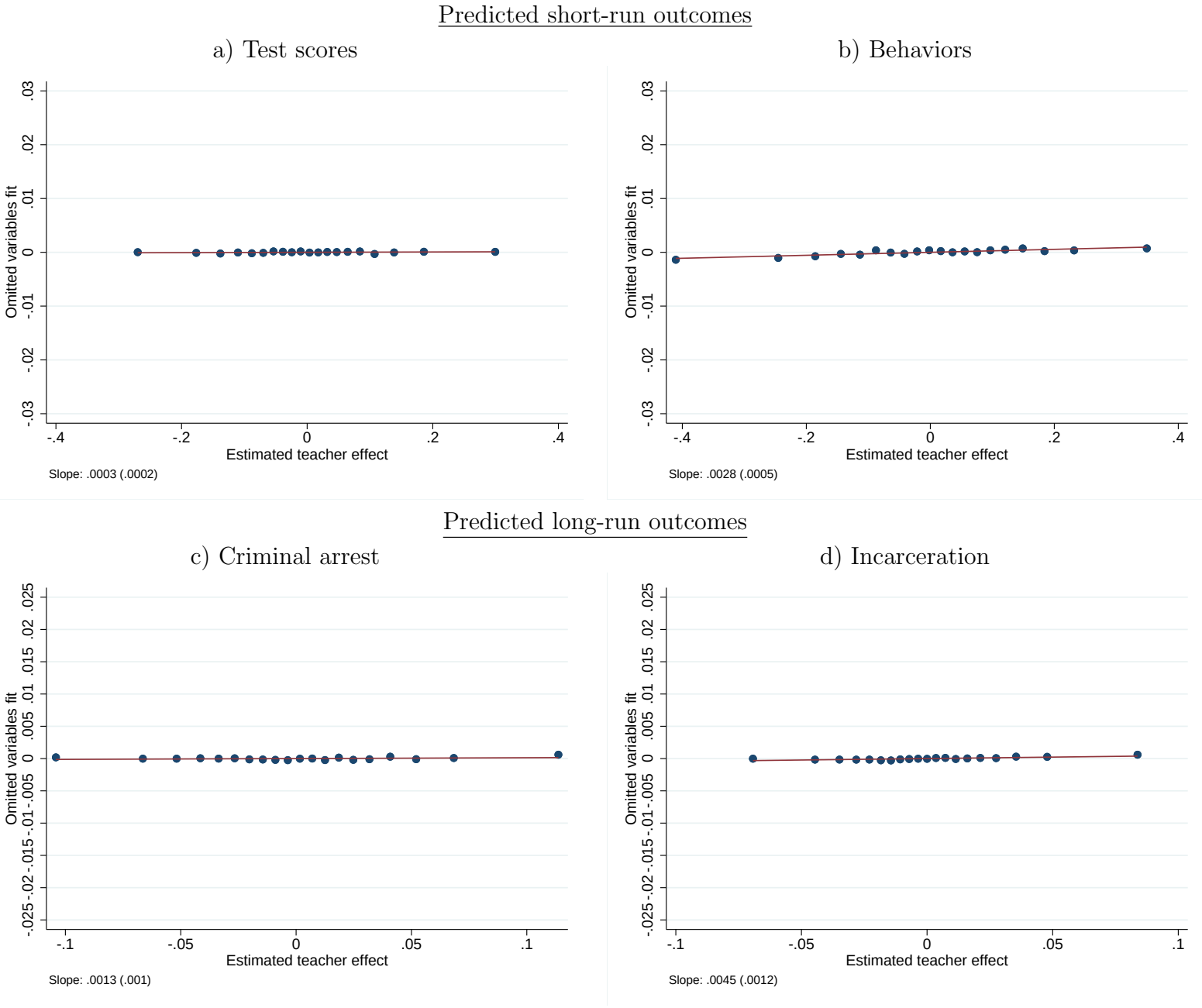
for improving students long-run outcomes. Moreover, policies based solely on teachers' test score quality may inadvertently remove teachers with important impacts on students' future CJC. Understanding these potentially trade-offs is essential for making effective policy in education.

Figure 1: Effects of teacher quality on long-run outcomes



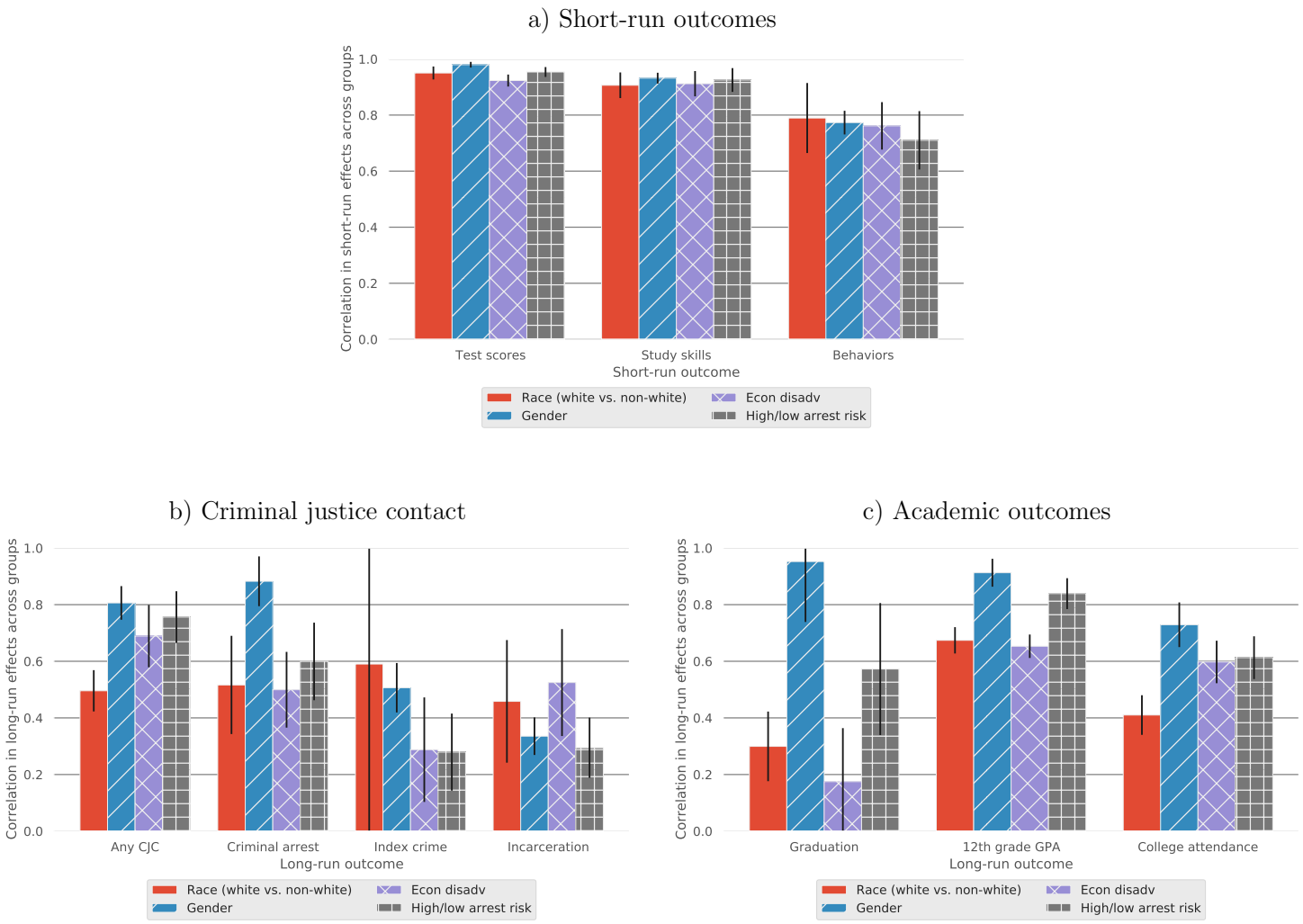
Notes: This figure presents the estimated effect of a 1 standard deviation in teacher quality as measured by short-run outcomes (x-axis) on long-run outcomes implied by estimates of the variance-covariance of teacher effects. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Numbers above/below each bar report effects as a percentage of the outcome mean. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 3.1 and use all available years for each outcome.

Figure 2: Assessing omitted variable bias in teacher effect estimates



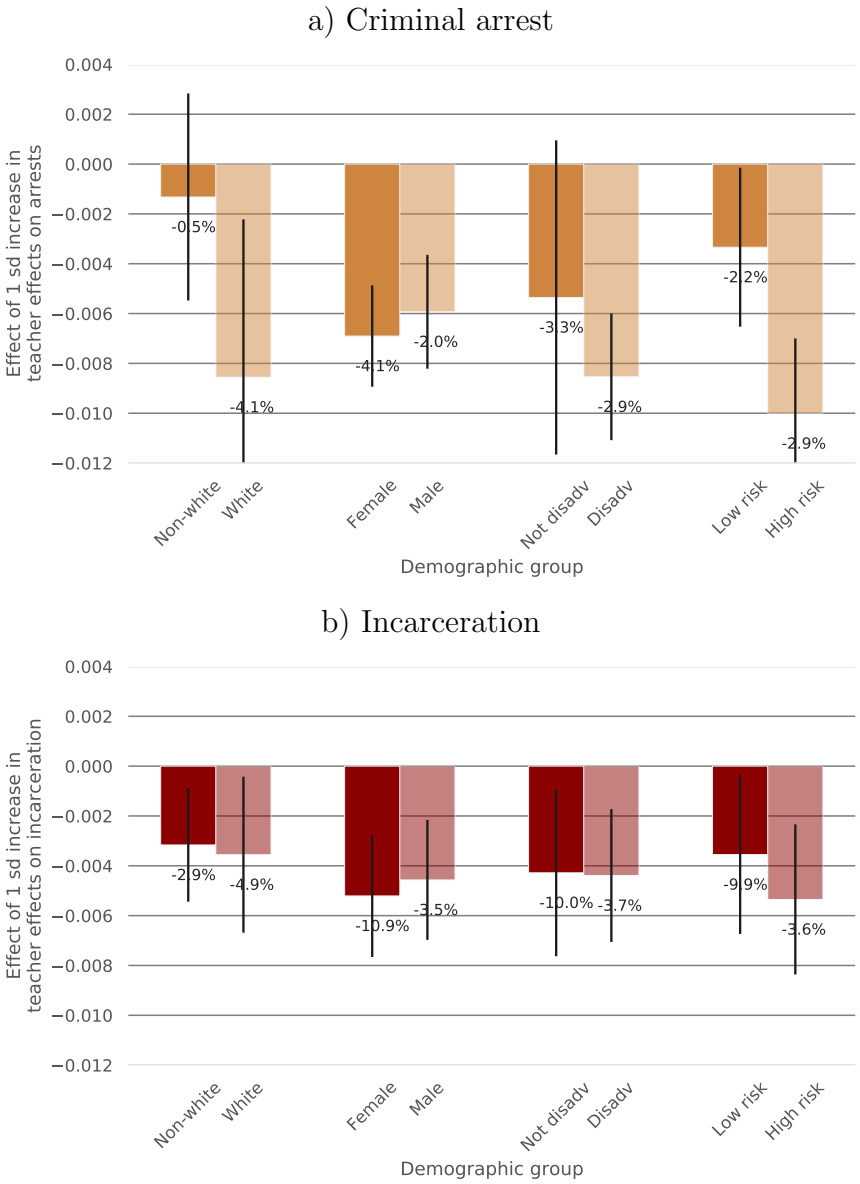
Notes: This figure presents a diagnostic test for whether the estimated teacher effects ($\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$ from Equation 2) are correlated with variables ($W'_{it}\hat{\rho}$ from Equation 9) that predict short- and long-run outcomes but were omitted when estimating the teacher effects. The flat slopes demonstrates that teacher effect estimates are insensitive to the inclusion of these omitted variables. Following Chetty, Friedman and Rockoff (2014*a*) we include parental education and twice lagged test scores among the omitted variables. We also include twins indicators as omitted variables, with all non-twins assigned to a separate indicator. Results change little when regressing $W'_{it}\hat{\rho}$ on $\hat{\alpha}_{it}$ in the sample of twins only. Teacher effect estimators include the full set of covariates described in Section 3.1 and use all available years for each outcome.

Figure 3: Correlation in teacher effects across groups



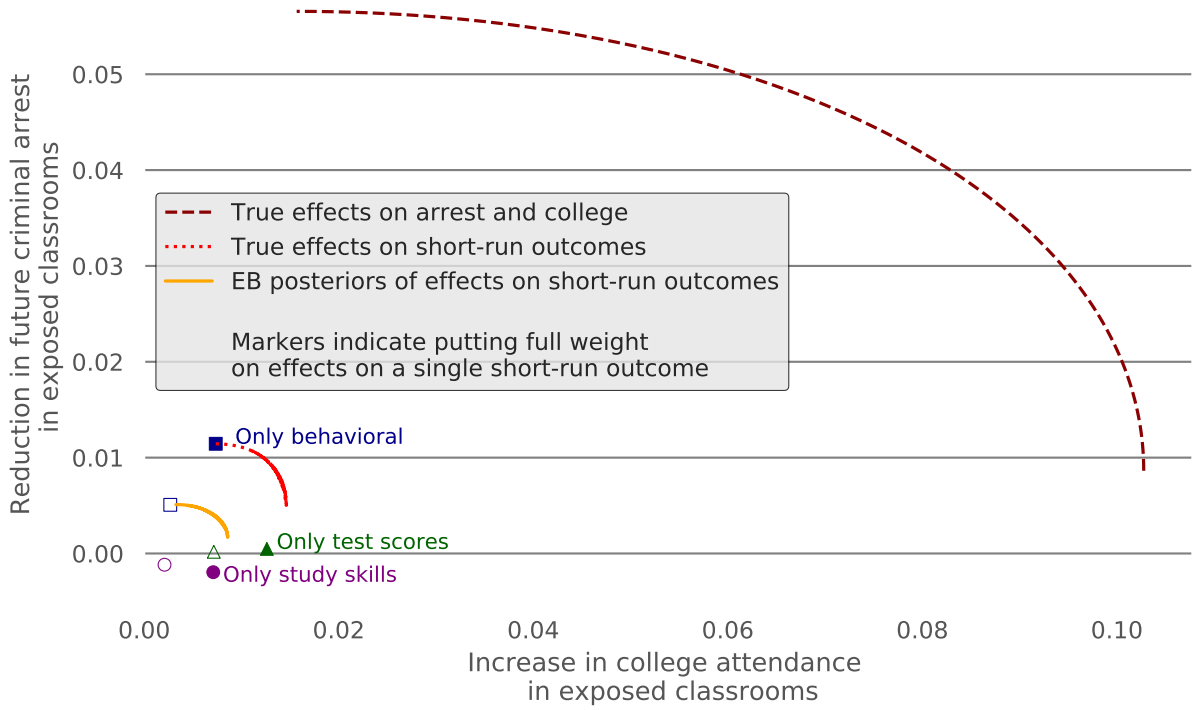
Notes: This figure presents the estimated correlation in teacher effects on short-run outcomes (panel a) and long-run outcomes (panels b and c) across groups of students. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 3.1 and use all available years for each outcome.

Figure 4: Heterogeneous impacts of exposure to teachers who improve behaviors



Notes: This figure presents the estimated effect of a one standard deviation in teacher quality as measured by impacts on students' behaviors on long-run outcomes across groups of students. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 3.1 and use all available years for each outcome.

Figure 5: Effects of teacher removal policies on exposed students



Notes: This figure presents simulations of the impacts of replacing the bottom five percent of teachers with an average teacher on college attendance and future criminal arrests. The rightmost dotted maroon lines in reflect the frontiers achievable if teachers' true long-run effects were directly observed and used to identify which teachers to replace. The dashed red line reflects possibilities if teachers true short-run effects on test scores, behaviors, and study skills were observed and used to select teachers. The leftmost solid line shows possibilities using EB estimates of teacher effects on short-run outcomes instead. The markers indicate gains when putting full weight on a single short-run outcome to select teachers for replacement, with solid markers using true effects and hollow markers using EB posteriors. Teacher effect estimators include the full set of covariates described in Section 3.1 and use all available years for each outcome. All simulations assume teacher effects are jointly normally distributed.

Table 1: Summary statistics

	Full sample		Sample for which we observe CJC		Youth with a criminal arrest	
	Mean (1)	SD (2)	Mean (3)	SD (4)	Mean (5)	SD (6)
Demographics						
Male	0.50	0.50	0.50	0.50	0.64	0.48
Black	0.25	0.43	0.27	0.44	0.37	0.48
Economically disadvantged	0.58	0.49	0.61	0.49	0.73	0.44
Limited English	0.043	0.20	0.055	0.23	0.036	0.19
Parents have HS education or less	0.40	0.49	0.43	0.49	0.53	0.50
Parents have some college	0.45	0.50	0.48	0.50	0.37	0.48
Parents have 4-year degree	0.22	0.41	0.24	0.43	0.15	0.36
Short-run outcomes						
Standardized reading scores	0.046	0.97	0.054	0.96	-0.23	0.94
Standardized math scores	0.061	0.97	0.066	0.97	-0.21	0.92
Days absent	9.11	9.20	9.08	9.44	10.9	11.1
Any discipline	0.17	0.37	0.17	0.38	0.31	0.46
Any out-of-school suspension	0.080	0.27	0.095	0.29	0.20	0.40
Repeat grade	0.0088	0.093	0.0087	0.093	0.015	0.12
Behavioral index	0	1.10	-0.025	1.12	0.44	1.37
Time spent on homework	0.023	0.99	0.023	0.99	-0.081	1.01
Time spent reading	0.0052	0.99	0.0041	0.99	-0.14	0.96
Time spent watching TV	-0.0052	0.98	-0.0078	0.98	0.085	1.02
Study skills index	0	1.09	0.0054	1.09	-0.17	1.09
Long-run outcomes						
12th grade GPA (0-6 scale)	3.13	0.95	3.12	0.95	2.64	0.87
12th grade class rank	0.48	0.29	0.48	0.28	0.61	0.26
Graduate high school	0.91	0.28	0.92	0.28	0.81	0.40
Plans to attend 4-year college	0.46	0.50	0.46	0.50	0.33	0.47
Any CJC 16-21	0.44	0.50	0.44	0.50	1	0
Traffic infraction	0.33	0.47	0.33	0.47	0.63	0.48
Criminal arrest	0.24	0.43	0.24	0.43	1	0
Index crime arrest	0.10	0.31	0.10	0.31	0.44	0.50
Criminal conviction	0.10	0.30	0.10	0.30	0.43	0.49
Incarcerated	0.089	0.29	0.089	0.29	0.36	0.48
N student-subject-years	9779708		4159500		984349	
N teachers	39707		27236		27202	
N students	1953547		755457		179484	
N twin pairs	18213		12516		4149	

Notes: This table presents summary statistics for demographic characteristics, short-run outcomes, and long-run outcomes for the analysis sample, the sample of students for which we observe CJC outcomes, and a sub-sample of students with a criminal arrest between ages 16 to 21. Not all outcomes are observed in all years; summary statistics reflect means and standard deviations for non-missing data only. In each analysis, we use the largest sample possible given when an outcome is studied. See Section 2 for additional details on data construction and outcome coverage by year. Note that the sample of youth with an arrest drops individuals for whom CJC outcomes are unobserved.

Table 2: Direct effects on long-run outcomes

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	College attendance	Graduation
Any CJC	0.035 (0.0000)	0.779 (0.0196)	0.513 (0.0371)	0.479 (0.0279)	-0.055 (0.0264)	-0.103 (0.0368)	-0.162 (0.0348)
Criminal arrest		0.027 (0.0000)	0.816 (0.0352)	0.583 (0.0244)	-0.150 (0.0660)	-0.153 (0.0468)	-0.262 (0.0431)
Index crime			0.018 (0.0000)	0.507 (0.0257)	-0.126 (0.1193)	-0.064 (0.0491)	-0.352 (0.0621)
Incarceration				0.021 (0.0000)	-0.101 (0.0719)	-0.175 (0.0609)	-0.281 (0.0476)
12th grade GPA					0.116 (0.0006)	0.394 (0.0469)	0.312 (0.1664)
College attendance						0.050 (0.0002)	0.256 (0.0616)
Graduation							0.023 (0.0001)

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects on long-run outcomes. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Any CJC refers to any interaction recorded in the criminal justice records between the ages of 16 and 21 inclusive. Criminal arrest excludes non-criminal interactions (e.g., traffic infractions). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' reported plans to attend a four-year college reported after graduation. Graduation is an indicator for graduating high school. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome.

Table 3: Teacher effects on short-run outcomes

	Test scores	Math scores	Reading scores	Study skills	Behaviors
Test scores	0.121 (0.0001)	0.909 (0.0025)	0.810 (0.0071)	0.317 (0.0089)	0.056 (0.0100)
Math scores		0.134 (0.0001)	0.675 (0.0076)	0.279 (0.0076)	0.047 (0.0101)
Reading scores			0.073 (0.0001)	0.337 (0.0210)	0.071 (0.0122)
Study skills				0.183 (0.0007)	0.033 (0.0132)
Behaviors					0.125 (0.0004)

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects on short-run outcomes. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading.

Table 4: Implied regression of long-run effects on short-run effects

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	-0.001 (0.005)	-0.003 (0.004)	-0.003 (0.003)	0.000 (0.003)	0.097 (0.012)	0.010 (0.003)	0.045 (0.008)
Behaviors	-0.059 (0.007)	-0.048 (0.006)	-0.038 (0.004)	-0.023 (0.004)	0.124 (0.016)	0.039 (0.004)	0.029 (0.010)
Study skills	-0.004 (0.004)	0.007 (0.004)	0.002 (0.003)	-0.009 (0.003)	-0.014 (0.011)	0.001 (0.003)	0.009 (0.008)
$sd(\alpha_j^y)$	0.035 (0.000)	0.027 (0.000)	0.018 (0.000)	0.021 (0.000)	0.116 (0.001)	0.023 (0.000)	0.050 (0.000)
R^2	0.039	0.042	0.059	0.023	0.025	0.041	0.020

Notes: This table presents the coefficients from a regression of long-run outcomes on short-run teacher effects implied by variance-covariance matrix of short- and long-run teachers effects. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. The final two rows report estimated standard deviations of teacher effects on the long-run outcome and the R^2 from the regression. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome.

Table 5: Instrumental variables tests for forecast unbiased teacher effects

(a) Short-run outcomes

	Test scores		Behaviors		Study skills	
	(1) Schl-grd	(2) Schl	(3) Schl-grd	(4) Schl	(5) Schl-grd	(6) Schl
$\hat{\alpha}_j$	1.002	1.052	0.938	0.984	1.043	1.087
	(0.0120)	(0.0157)	(0.0678)	(0.117)	(0.0353)	(0.0595)
Observations	9779708	9779708	5422682	5422682	3404657	3404657
R^2	0.753	0.752	0.247	0.244	0.180	0.176
Design controls	✓	✓	✓	✓	✓	✓
First stage F	51914	32164	3349	1205	6865	2557
P-value for $H_0 : \lambda = 1$	.873	.001	.363	.891	.229	.145

(b) Long-run outcomes

	Criminal arrest		Incarceration		College bound	
	(1) Schl-grd	(2) Schl	(3) Schl-grd	(4) Schl	(5) Schl-grd	(6) Schl
$\hat{\alpha}_j$	1.090	1.190	1.156	1.246	0.983	0.807
	(0.0824)	(0.311)	(0.0890)	(0.319)	(0.0535)	(0.167)
Observations	4159500	4159500	4159500	4159500	3205422	3205422
R^2	0.0916	0.0899	0.0836	0.0816	0.282	0.279
Design controls	✓	✓	✓	✓	✓	✓
First stage F	4825	386	4032	397	4399	651
P-value for $H_0 : \lambda = 1$	.276	.543	.079	.44	.745	.248

Notes: This table presents instrumental variable tests for bias in estimated teacher effects on short- and long-run outcomes, where an estimate of 1 implies forecast unbiased estimates. Design controls include the full set of covariates described in Section 3.1 and using all available years for each outcome. The reported coefficient on $\hat{\alpha}_{it}$ is estimated via 2SLS using a teacher switching instrument defined at the school-grade (odd columns) or school-level (even columns). The instrument is the product of an indicator for new teacher entry into student i's school-grade or school at time t times the mean of $\hat{\alpha}_j$ for all entering teachers estimated in all other school-grades or schools. Only entries where at least one new teacher's effects are estimable in other schools or school grades are included in the instrument. Means are weighted by number of students assigned at time t. All regressions include an indicator for any teacher entry. Standard errors clustered at the student level are reported in parentheses.

Table 6: Heterogeneity in teacher effects on CJC

	Race		Sex		Econ. disadvantaged		Arrest risk	
	(1) White	(2) Non-White	(3) Boys	(4) Girls	(5) Yes	(6) No	(7) High	(8) Low
Criminal arrest	0.0247	0.0358	0.0298	0.0311	0.0272	0.0297	0.0342	0.0263
	(0.000135)	(0.000134)	(0.000214)	(0.000109)	(0.0000622)	(0.0000846)	(0.000211)	(0.000163)
Any CJC	0.0391	0.0370	0.0412	0.0364	0.0364	0.0367	0.0443	0.0436
	(0.000153)	(0.000136)	(0.000213)	(0.000117)	(0.0000958)	(0.0000981)	(0.000200)	(0.000155)
Incarceration	0.0172	0.0308	0.0156	0.0270	0.0211	0.0303	0.0320	0.0201
	(0.0000920)	(0.0000874)	(0.000160)	(0.0000733)	(0.0000294)	(0.0000618)	(0.000182)	(0.000153)
Index crime	0.0177	0.0248	0.0147	0.0239	0.0212	0.0209	0.0276	0.0113
	(0.000103)	(0.000100)	(0.000184)	(0.0000865)	(0.0000410)	(0.0000530)	(0.000187)	(0.000157)

Notes: This table presents estimates of the standard deviations of teacher effects on future CJC across various sub-populations: white vs. non-white (Columns 1 and 2), boys vs. girls (Columns 3 and 4), economically disadvantaged vs. not (Columns 5 and 6), and students with high vs. low predicted risk of a future arrest (Columns 7 and 8). Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). Traffic citations includes only non-criminal traffic violations. Index crimes includes arrests for Uniform Crime Reporting index crimes: aggravated assault, forcible rape, murder, robbery, arson, burglary, larceny/theft, and motor vehicle theft. Incarceration refers to any incarceration sentence in local jails or state prisons. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, 25: 95–135.

Abaluck, Jason, Mauricio Caceres Bravo, Peter Hull, and Amanda Starc. 2020. "Mortality Effects and Choice Across Private Health Insurance Plans." *Working Paper.*

Abdulkadiroğlu, Atila, Parag A Pathak, Jonathan Schellenberg, and Christopher R Walters. 2020. "Do parents value school effectiveness?" *American Economic Review*, 110(5): 1502–39.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *The Quarterly Journal of Economics*, 132(2): 871–919.

Bacher-Hicks, Andrew, Stephen B. Billings, and David J. Deming. 2019. "The School to Prison Pipeline: Long-Run Impacts of School Suspensions on Adult Crime." National Bureau of Economic Research Working Paper 26257.

Bacher-Hicks, Andrew, Thomas J Kane, and Douglas O Staiger. 2014*a*. "Validating teacher effect estimates using changes in teacher assignments in Los Angeles." National Bureau of Economic Research.

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. 2014*b*. "Validating Teacher Effects Estimates Using Changes in Teacher Assignments in Los Angeles." *NBER Working Paper No. 20657.*

Backes, Ben, James Cowan, Dan Goldhaber, and Roddy Theobald. 2022. "Teachers and Students' Postsecondary Outcomes: Testing the Predictive Power of Test and Nontest Teacher Quality Measures." CALDER Working Paper.

Bates, Michael D, Michael Dinerstein, Andrew C Johnston, and Isaac Sorkin. 2022. "Teacher Labor Market Equilibrium and Student Achievement." National Bureau of Economic Research.

Bau, Natalie. 2022. "Estimating an equilibrium model of horizontal competition in education." *Journal of Political Economy*, 130(7).

Bau, Natalie, and Jishnu Das. 2020. "Teacher value added in a low-income country." *American Economic Journal: Economic Policy*, 12(1): 62–96.

Bell, Brian, Rui Costa, and Stephen J Machin. 2018. "Why does education reduce crime?" CEPR Discussion Paper No. DP13162.

Bertrand, Marianne, and Jessica Pan. 2013. "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior." *American Economic Journal: Applied Economics*, 5(1): 32–64.

Beuermann, Diether, C Kirabo Jackson, Laia Navarro-Sola, and Francisco Pardo. 2018. "What is a good school, and can parents tell? Evidence on the multidimensionality of school output." National Bureau of Economic Research.

Biasi, Barbara. 2021. "The Labor Market for Teachers under Different Pay Schemes." *American Economic Journal: Economic Policy*, 13(3): 63–102.

Biasi, Barbara, Chao Fu, and John Stromme. 2021. "Equilibrium in the Market for Public School Teachers: District Wage Strategies and Teacher Comparative Advantage." National Bureau of Economic Research.

Billings, Stephen B., David J. Deming, and Jonah Rockoff. 2013. "School Segregation, Educational Attainment, and Crime: Evidence from the End of Busing in Charlotte-Mecklenburg." *The Quarterly Journal of Economics*, 129(1): 435–476.

Billings, Stephen B, David J Deming, and Stephen L Ross. 2019. "Partners in crime." *American Economic Journal: Applied Economics*, 11(1): 126–50.

Borghans, Lex, Baster Weel, and Bruce A. Weinberg. 2008. "Interpersonal Styles and Labor Market Outcomes." *Journal of Human Resources*, 43(4): 815–858.

Brame, Robert, Michael G. Turner, Raymond Paternoster, and Shawn D. Bushway. 2012. "Cumulative prevalence of arrest from ages 8 to 23 in a national sample." *Pediatrics*, 129(1): 21–27.

Brame, Robert, Shawn D. Bushway, Ray Paternoster, and Michael G. Turner. 2014. "Demographic Patterns of Cumulative Arrest Prevalence By Ages 18 and 23." *Crime and Delinquency*, 60(3): 471–486.

Bruhn, Jesse. 2020. "The consequences of sorting for understanding school quality." *Unpublished working paper.*

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014*a*. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104(9).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014*b*. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2016. "Using Lagged Outcomes to Evaluate Bias in Value-Added Models." *American Economic Review*, 106(5): 393–99.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2017. "Measuring the Impacts of Teachers: Reply." *American Economic Review*, 107(6): 1685–1717.

Condie, Scott, Lars Lefgren, and David Sims. 2014. "Teacher heterogeneity, value-added and education policy." *Economics of Education Review*, 40: 76–92.

Cook, Philip J., and Songman Kang. 2016. "Birthdays, Schooling, and Crime: Regression-Discontinuity Analysis of School Performance, Delinquency, Dropout, and Crime Initiation." *American Economic Journal: Applied Economics*, 8(1): 33–57.

Cullen, Julie Berry, Brian A Jacob, and Steven Levitt. 2006. "The Effect of School Choice on Participants: Evidence from Randomized Lotteries." *Econometrica*, 74(5): 1191–1230.

Cunha, Flavio, and James J Heckman. 2008. "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation." *Journal of human resources*, 43(4): 738–782.

Cunha, Flavio, James J Heckman, and Susanne M Schennach. 2010. "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78(3): 883–931.

Dee, Thomas S. 2005. "A Teacher like Me: Does Race, Ethnicity, or Gender Matter?" *The American Economic Review*, 95(2): 158–165.

Delgado, William. 2021. "Heterogeneous Teacher Effects, Comparative Advantage, and Match Quality." Working Paper.

Deming, David J. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics*, 1(3): 111–34.

Deming, David J. 2011. " Better Schools, Less Crime?" *The Quarterly Journal of Economics*, 126(4): 2063–2115.

Deming, David J. 2017. "The Growing Importance of Social Skills in the Labor Market." *The Quarterly Journal of Economics*, 132(4): 1593–1640.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. "Grit: perseverance and passion for long-term goals." *Journal of personality and social psychology*, 92: 1087–101.

Gershenson, Seth. 2016. "Linking teacher quality, student attendance, and student achievement." *Education Finance and Policy*, 11(2): 125–149.

Gershenson, Seth, Cassandra M. D Hart, Joshua Hyman, Constance Lindsay, and Nicholas W Papageorge. 2018. "The Long-Run Impacts of Same-Race Teachers." National Bureau of Economic Research Working Paper 25254.

Gilraine, Michael, Jiaying Gu, Robert McMillan, et al. 2021. "A Nonparametric Method for Estimating Teacher Value-Added."

Gray-Lobe, Guthrie, Parag A Pathak, and Christopher R Walters. 2021. "The Long-Term Effects of Universal Preschool in Boston." National Bureau of Economic Research.

Hanushek, Eric A. 2011. "The economic value of higher teacher quality." *Economics of Education review*, 30(3): 466–479.

Heckman, James J., and Tim Kautz. 2012. "Hard evidence on soft skills." *Labour Economics*, 19(4): 451–464.

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review*, 91(2): 145–149.

Heckman, James J, Jora Stixrud, and Sergio Urzua. 2006. "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior." *Journal of Labor Economics*, 24(3): 411–482.

Heckman, James J, Seong Hyeok Moon, Rodrigo Pinto, Peter A Savelyev, and Adam Yavitz. 2010a. "The rate of return to the HighScope Perry Preschool Program." *Journal of Public Economics*, 94(1): 114–128.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review*, 103(6): 2052–86.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack. 2016. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." *The Quarterly Journal of Economics*, 132(1): 1–54.

Jackson, C Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes." *Journal of Political Economy*, 126(5): 2072–2107.

Jackson, C. Kirabo, Shanette C. Porter, John Q. Easton, Alyssa Blanchard, and Sebastián Kiguel. 2020. "School Effects on Socioemotional Development, School-Based Arrests, and Educational Attainment." *American Economic Review: Insights*, 2(4): 491–508.

Kane, Thomas J, and Douglas O Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14607.

Karoui, Noureddine El, and Elizabeth Purdom. 2016. "Can we trust the bootstrap in high-dimension?" *arXiv preprint arXiv:1608.00696*.

Kline, Patrick M., Evan K. Rose, and Christopher R. Walters. 2021. "Systemic Discrimination Among Large U.S. Employers." National Bureau of Economic Research.

Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten. 2020. "Leave-Out Estimation of Variance Components." *Econometrica*, 88(5): 1859–1898.

Krueger, Alan B, and Lawrence H Summers. 1988. "Efficiency wages and the inter-industry wage structure." *Econometrica: Journal of the Econometric Society*, 259–293.

Lindqvist, Erik, and Roine Vestman. 2011. "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." *American Economic Journal: Applied Economics*, 3(1): 101–28.

Lleras, Christy. 2008. "Do skills and behaviors in high school matter? The contribution of noncognitive factors in explaining differences in educational attainment and earnings." *Social Science Research*, 37(3): 888–902.

Lochner, Lance. 2011. "Nonproduction Benefits of Education: Crime, Health, and Good Citizenship." In *Handbook of the Economics of Education.* , ed. E. Hanushek, S. Machin and L. Woessmann.

Lochner, Lance, and Enrico Moretti. 2004. "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports." *American Economic Review*, 94(1): 155–189.

Neal, Derek. 2011. "The design of performance pay in education." In *Handbook of the Economics of Education.* Vol. 4, 495–550.

Papp, Jordan, and Michael Mueller-Smith. 2021. "Benchmarking the Criminal Justice Administrative Records System's Data Infrastructure." University of Michigan Working Paper.

Petek, Nathan, and Nolan Pope. 2021. "The multidimensional impact of teachers on students." Working Paper.

Reynolds, Arthur J., Judy A. Temple, and Suh-Ruu Ou. 2010. "Preschool education, educational attainment, and crime prevention: Contributions of cognitive and non-cognitive skills." *Children and Youth Services Review*, 32(8): 1054–1063.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417–458.

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

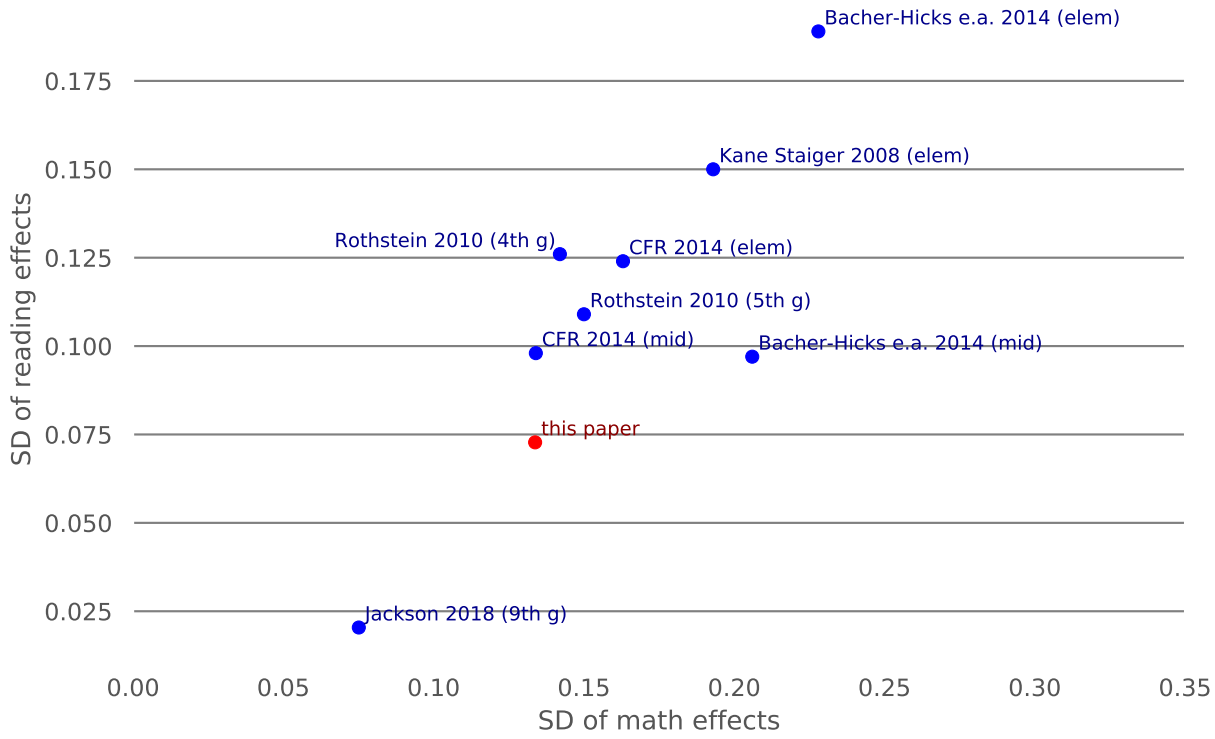
Rothstein, Jesse. 2017. "Measuring the impacts of teachers: Comment." *American Economic Review*, 107(6): 1656–84.

Sorensen, Lucy C., Shawn D. Bushway, and Elizabeth J. Gifford. 2019. "Getting tough? The effects of discretionary principal discipline on student outcomes." *Education Finance and Policy*, 1–74.

Waddell, Gen R. 2006. "Labor-market consequences of poor attitude and low self-esteem in youth." *Economic Inquiry*, 44(1): 69–97.

A Additional figures and tables

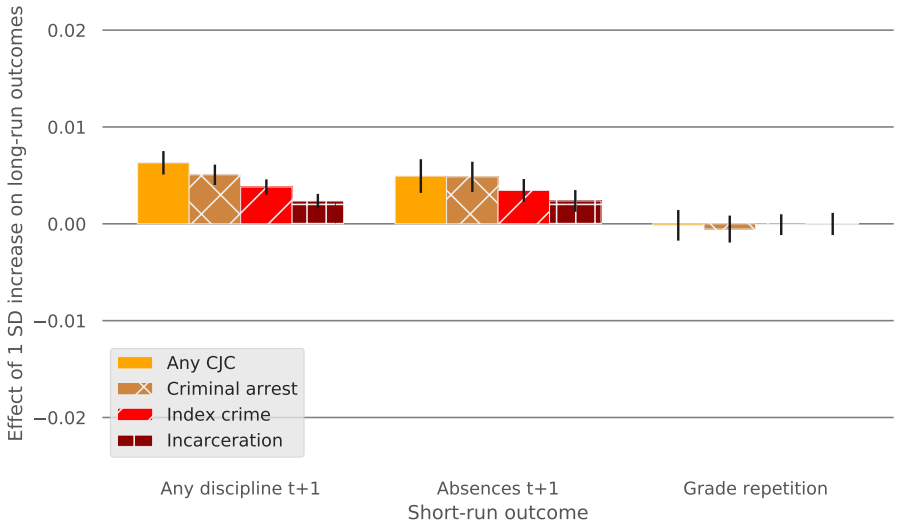
Figure A.1: Teacher test score effects in the literature



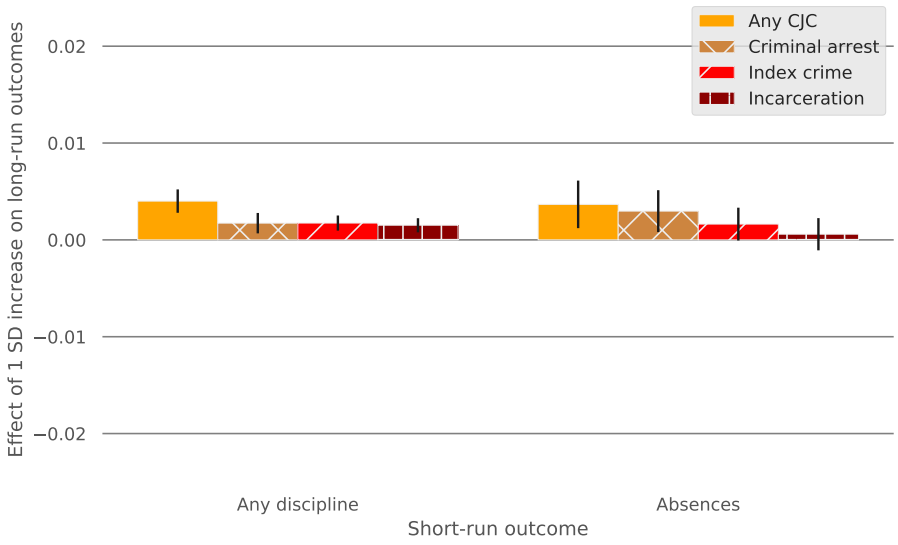
Notes: This figure compares estimated standard deviation of teacher effects on math and reading scores to comparable estimates in the literature. "Mid" indicates estimates for middle school students and "elem" indicates elementary school students. Our estimates straddle those from studies that focus on elementary students vs. those that focus on older students (e.g., Jackson (2018)).

Figure A.2: Long-run effects of specific behavioral quality dimensions

(a) Behavioral measures at $t+1$



(b) Contemporaneous behavioral measures

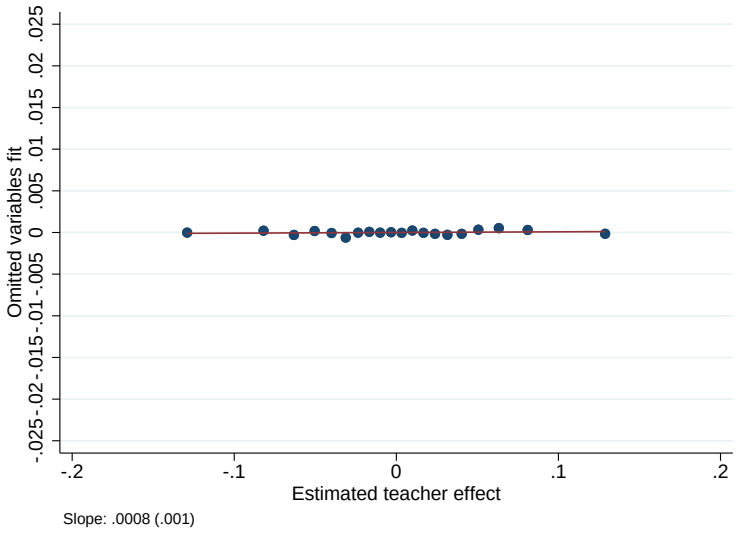


Notes: This figure presents the estimated effect of a one standard deviation in teacher quality as measured by short-run outcomes (x-axis) on long-run outcomes implied by estimates of the variance-covariance of teacher effects. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Any discipline $t+1$ is an indicator for any discipline, including in- and out-of-school suspensions, the year after the student and teacher shared a classroom. Absences is the number of days absent in the year after assignment. Grade repetition is an indicator for repeating the current grade. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 3.1 and use all available years for each outcome.

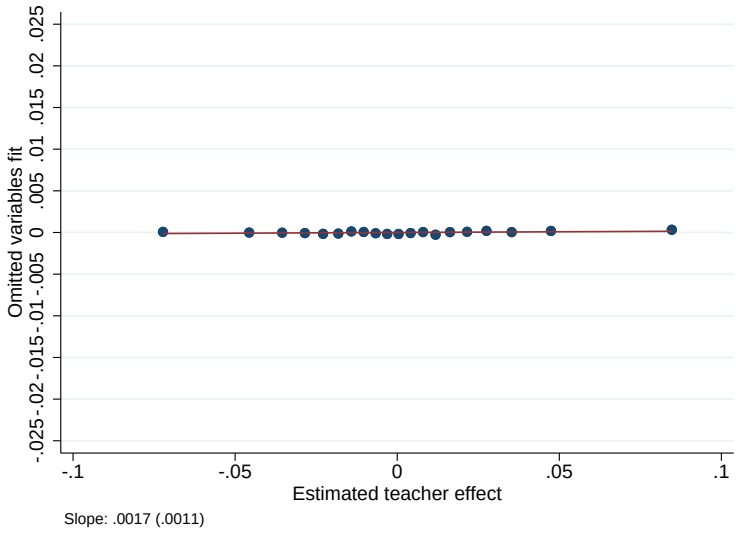
Figure A.3: Assessing omitted variable bias in additional long-run outcomes

Predicted long-run CJC outcomes

a) Any CJC

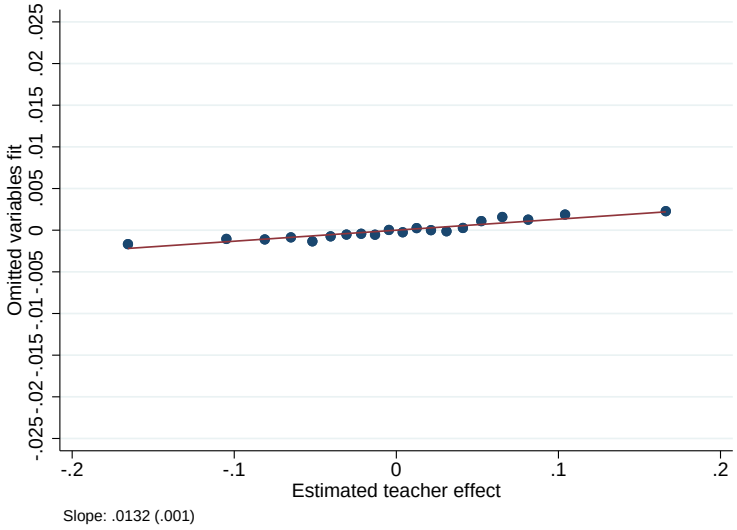


b) Arrest for index crime

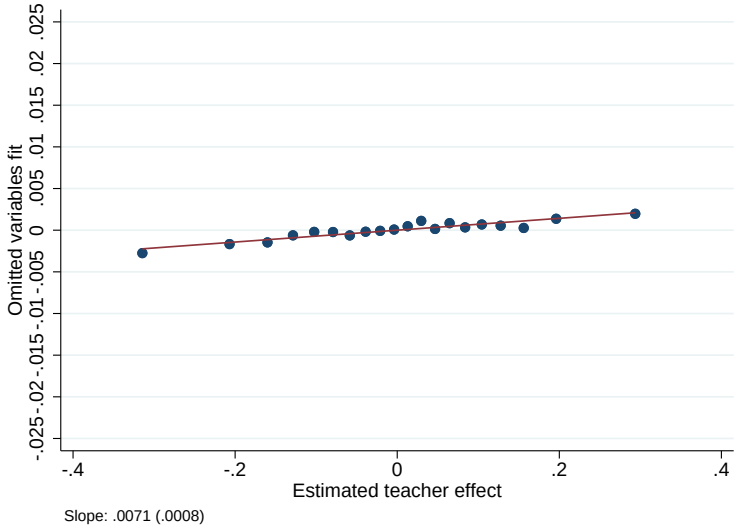


Predicted long-run academic outcomes

c) College bound

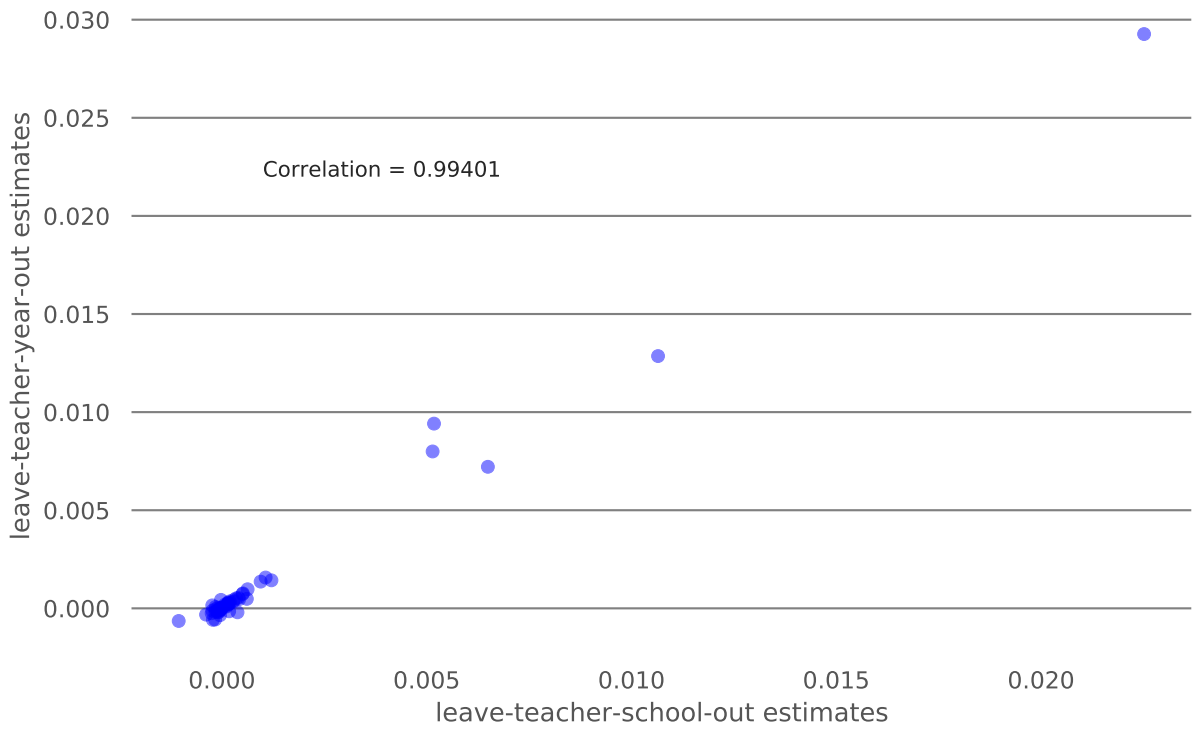


d) 12th grade GPA



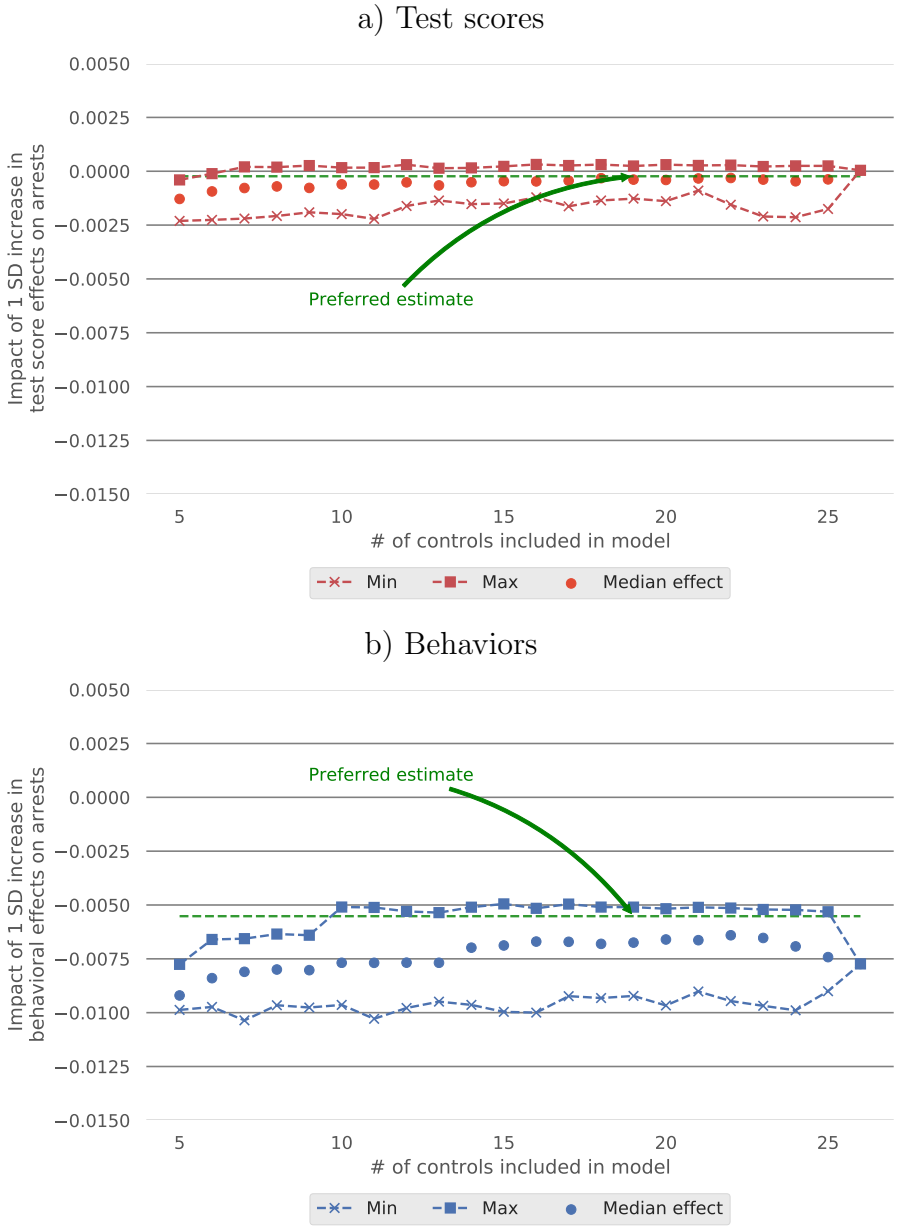
Notes: This figure presents a diagnostic test for whether the estimated teacher effects ($\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$ from Equation 2) are correlated with predictions based on omitted variables ($W'_{it}\hat{\rho}$ from Equation 9) that are predictable of the short- and long-run outcomes but have not been used when estimating the teacher effects. Following Chetty, Friedman and Rockoff (2014*a*) we include parental education and twice lagged test scores among the omitted variables. We also include twins indicators as omitted variables, with all non-twins assigned to a separate indicator. Results change little when regressing $W'_{it}\hat{\rho}$ on $\hat{\alpha}_{it}$ in the sample of twins only. Teacher effect estimators include the full set of covariates described in Section 3.1 and use all available years for each outcome.

Figure A.4: Correlation between primary and between-school estimates



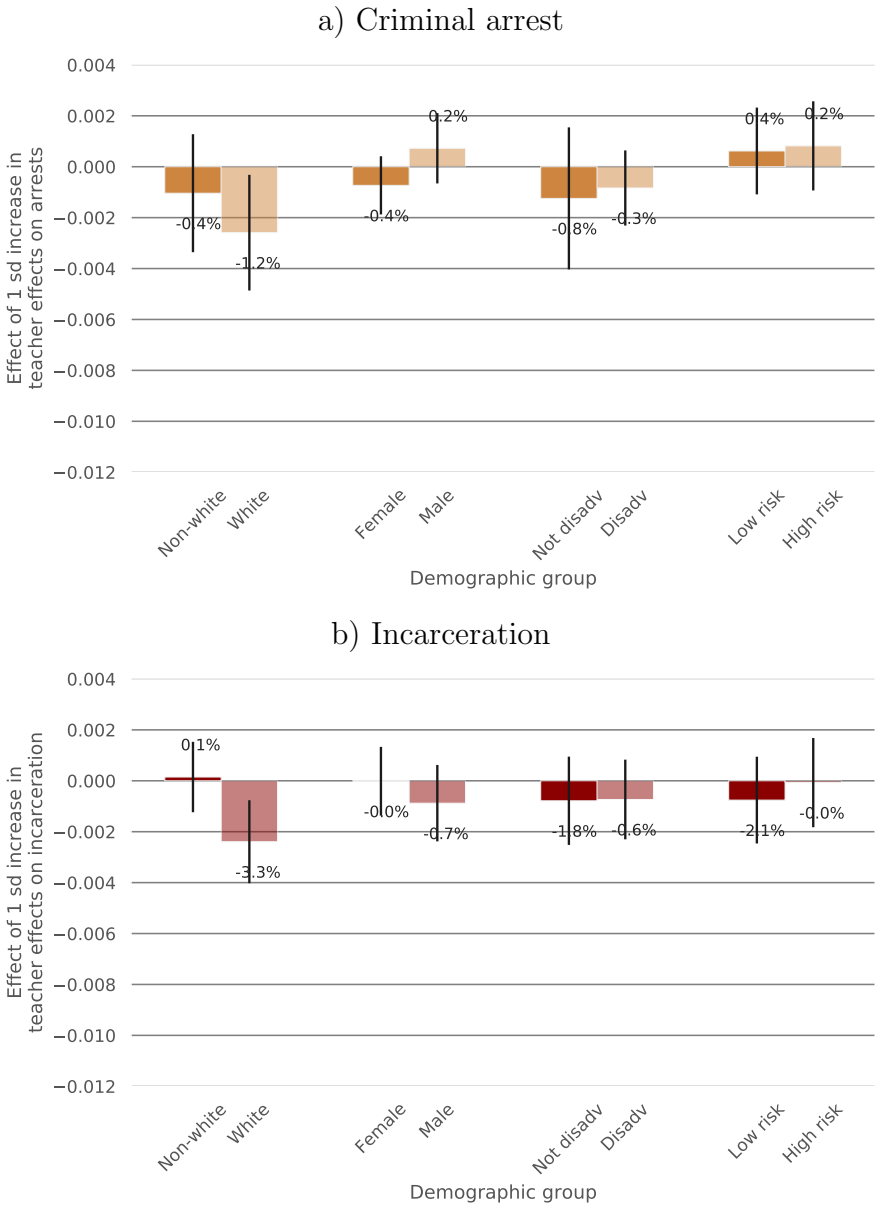
Notes: This figures shows the relationship between variance estimates of teacher effects using leave-year-out (Equation 4) and leave-school-out (Equation 11) estimators. Each point in the figure is a teacher effect variance estimate for a different outcome. The x-axis reports the value of the estimated variance of teacher effects using leave-year-out estimators. The y-axis reports the value of the estimate when using leave-school-out estimators.

Figure A.5: Specification sensitivity of teacher quality impacts on arrests



Notes: This figures shows the specification sensitivity of estimated effects of one standard deviation increase in teacher quality on future criminal arrests. We estimate the variance-covariance of teacher effects from 811 different models that vary the number of included controls. All models include lag third-degree polynomials in math and reading scores interacted with grade, and year-grade-subject FEs. The x-axis shows the quantity of other controls included from among school, school-grade-year, or school-grade-classroom-year means of other included covariates, lag absences and discipline, educational and behavioral special needs, and academically gifted indicators, limited English proficiency status, gender and race, parental education, grade repetition, and twice-lagged scores. The graph reports the min, median, and max effect estimate among models with the same number of controls.

Figure A.6: Heterogeneous impacts of exposure to teachers who improve test scores



Notes: This figure presents the estimated effect of a one standard deviation in teacher quality as measured by impacts on students' test scores on long-run outcomes across groups of students. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 3.1 and use all available years for each outcome.

Table A.1: Teacher effects on short-run behaviors

	Any discipline $t+1$	Any OSS $t+1$	Grade repetition	Days absent $t+1$
Any discipline $t+1$	0.059 (0.0001)	0.510 (0.0217)	0.047 (0.0291)	0.085 (0.3167)
Any OSS $t+1$		0.026 (0.0000)	0.038 (0.0262)	0.186 (3.6149)
Grade repetition			0.008 (0.0000)	-0.020 (4.4779)
Days absent $t+1$				0.759 (0.0254)

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects on behavioral proxies for non-cognitive skills. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Any discipline refers to any detention, in-school suspension, out of school suspension, or other disciplinary event recorded in the year. Any OSS refers to any out of school suspension. $t+1$ indicates the event occurred the year following (e.g., in 5th grade for 4th grade students). Grade repetition refers to repeating the grade at time t. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome.

Table A.2: Correlation between short- and long-run teacher effects

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	College attendance	Graduation
Test scores	-0.022 (0.0145)	-0.008 (0.0155)	-0.023 (0.0188)	-0.030 (0.0157)	0.100 (0.0099)	0.123 (0.0145)	0.065 (0.0143)
Behaviors	-0.199 (0.0455)	-0.202 (0.0665)	-0.242 (0.1642)	-0.126 (0.0669)	0.126 (0.0140)	0.072 (0.0219)	0.198 (0.0765)
Study skills	-0.029 (0.0225)	0.035 (0.0240)	0.012 (0.0256)	-0.084 (0.0682)	0.013 (0.0155)	0.069 (0.0237)	0.027 (0.0262)

Notes: This table presents estimated correlation between teachers short- and long-run effects. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Any CJC refers to any interaction recorded in the criminal justice records between the ages of 16 and 21 inclusive. Criminal arrest excludes non-criminal interactions (e.g., traffic infractions). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' reported plans to attend a four-year college reported after graduation. Graduation is an indicator for graduating high school. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome.

Table A.3: Regression based estimates of teacher test score effects on long-run outcomes

	CJC outcomes				Academic outcomes		
	(1) Any CJC	(2) Criminal arrest	(3) Index crime	(4) Incarceration	(5) 12th grade GPA	(6) Graduation	(7) College bound
Test score VA	-0.0108	-0.00795	-0.00591	-0.00553	0.0781	0.0104	0.0368
	(0.00398)	(0.00332)	(0.00228)	(0.00232)	(0.00907)	(0.00220)	(0.00478)
Design controls	✓	✓	✓	✓	✓	✓	✓
1SD effect	-.0012	-.0008	-.0006	-.0006	.0083	.0011	.0039
R2	0.0480	0.0754	0.0596	0.0663	0.542	0.106	0.253
Observations	4159500	4159500	4159500	4159500	3429388	4623602	3205422

Notes: This table presents regressions of teacher test score value added calculated using the method in Chetty, Friedman and Rockoff (2014*a*) on long-run outcomes. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. This method allows for drift in teacher effects and accounts for measurement error by forming the best linear predictor of teacher effects in year t based on their impacts in all other years. The final row of the table presents the regression coefficient implied by our procedure. Teacher value-added estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome. Standard errors clustered at the student level are reported in parentheses.

Table A.4: Regression based estimates of teacher behavioral effects on long-run outcomes

	CJC outcomes				Academic outcomes		
	(1) Any CJC	(2) Criminal arrest	(3) Index crime	(4) Incarceration	(5) 12th grade GPA	(6) Graduation	(7) College bound
Behavioral index VA	-0.0517	-0.0453	-0.0354	-0.0233	0.179	0.0259	0.0361
	(0.00639)	(0.00541)	(0.00376)	(0.00365)	(0.0130)	(0.00346)	(0.00753)
Design controls	✓	✓	✓	✓	✓	✓	✓
1SD effect	-.0042	-.0036	-.0028	-.0019	.0142	.002	.0029
R2	0.0481	0.0751	0.0596	0.0651	0.543	0.106	0.256
Observations	3700227	3700227	3700227	3700227	2969019	4001899	2850488

Notes: This table presents regressions of teacher behavioral value added calculated using the method in Chetty, Friedman and Rockoff (2014*a*) on long-run outcomes. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. This method allows for drift in teacher effects and accounts for measurement error by forming the best linear predictor of teacher effects in year t based on their impacts in all other years. The final row of the table presents the regression coefficient implied by our procedure. Teacher value-added estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome. Standard errors clustered at the student level are reported in parentheses.

Table A.5: Omitted variables bias tests for short-run teacher effects

	Test scores		Behvioral index		Study skills index	
	(1)	(2)	(3)	(4)	(5)	(6)
	Y	$\hat{Y}$	Y	$\hat{Y}$	Y	$\hat{Y}$
No high school	-0.0491		-0.102		-0.0533	
	(0.000932)		(0.00268)		(0.00255)	
High school only	-0.0329		-0.0768		-0.0239	
	(0.000780)		(0.00187)		(0.00239)	
Some college	0.0157		0.0476		0.0439	
	(0.000811)		(0.00204)		(0.00244)	
BA or more	0.0260		0.0327		0.0668	
	(0.000643)		(0.00191)		(0.00170)	
Lag 2 math	0.114		0.0311		0.00129	
	(0.000401)		(0.000974)		(0.00137)	
Lag 2 reading	0.126		0.0144		0.0582	
	(0.000384)		(0.000928)		(0.00129)	
Teacher effect		0.000309		0.00277		0.00202
		(0.000187)		(0.000452)		(0.000490)
Observations	9757562	9757562	5415717	5415717	3387386	3387386
R2	0.761	0.770	0.252	0.872	0.189	0.584
Original R2	.7497		.2408		.1723	
Design controls	✓	✓	✓	✓	✓	✓
Twin FE	✓		✓		✓	

Notes: This table presents tests for omitted variable bias in estimated teacher effects on short-run outcomes. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. The odd columns regress the outcome listed in the sub-header on the excluded covariates, teacher dummies, and the full set of design controls described in Section 3.1. The evem columns regress predicted outcomes based on the excluded covariates on estimated teacher effects $\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$. Education variables refer to students' reported parental education, with an indicator for missing parental education data serving as the omitted category. Lag 2 math and reading refer to twice-lagged standardized test scores, with indicators for missing twice-lag scores included but not reported. Twin-effects include fixed effects for all twin pairs and an indicator for non-twin interacted with year. Results change little when regressing $\hat{Y}_{it}$ on $\hat{\alpha}_{it}$ in the sample of twins only. Original R^2 refers the R^2 of the regression without excluded covariates used to estimate teacher effects. Standard errors clustered at the student level are reported in parentheses.

Table A.6: Omitted variables bias tests for long-run teacher effects on outcomes related to criminal justice involvement

	Any arrest		Criminal arrest		Index crime		Incarceration	
	(1) Y	(2) $\hat{Y}$	(3) Y	(4) $\hat{Y}$	(5) Y	(6) $\hat{Y}$	(7) Y	(8) $\hat{Y}$
No high school	0.00300		0.0177		0.0198		0.0270	
	(0.00113)		(0.000954)		(0.000694)		(0.000642)	
High school only	-0.0000113		0.0184		0.0147		0.0173	
	(0.000866)		(0.000730)		(0.000531)		(0.000492)	
Some college	-0.00307		-0.0170		-0.0150		-0.0175	
	(0.000913)		(0.000770)		(0.000560)		(0.000518)	
BA or more	-0.0196		-0.0178		-0.00873		-0.00615	
	(0.000781)		(0.000658)		(0.000479)		(0.000443)	
Lag 2 math	0.0107		0.00268		-0.000369		0.000680	
	(0.000562)		(0.000474)		(0.000345)		(0.000319)	
Lag 2 reading	-0.00953		-0.00776		-0.00544		-0.00499	
	(0.000526)		(0.000443)		(0.000322)		(0.000298)	
Teacher effect		0.000794		0.00128		0.00169		0.00452
		(0.00101)		(0.00103)		(0.00115)		(0.00119)
Observations	4145532	4145532	4145532	4145532	4145532	4145532	4145532	4145532
R2	0.0811	0.718	0.107	0.702	0.0914	0.790	0.101	0.737
Original R2	.062		.0888		.0726		.0807	
Design controls	✓	✓	✓	✓	✓	✓	✓	✓
Twin FE	✓		✓		✓		✓	

Notes: This table presents tests for omitted variable bias in estimated teacher effects on long-run outcomes. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. The odd columns regress the outcome listed in the sub-header on the excluded covariates, teacher dummies, and the full set of design controls described in Section 3.1. The even columns regress predicted outcomes based on the excluded covariates on estimated teacher effects $\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$. Education variables refer to students' reported parental education, with an indicator for missing parental education data serving as the omitted category. Lag 2 math and reading refer to twice-lagged standardized test scores, with indicators for missing twice-lag scores included but not reported. Twin-effects include fixed effects for all twin pairs and an indicator for non-twin interacted with year. Results change little when regressing $\hat{Y}_{it}$ on $\hat{\alpha}_{it}$ in the sample of twins only. Original R^2 refers the R^2 of the regression without excluded covariates used to estimate teacher effects. Standard errors clustered at the student level are reported in parentheses.

Table A.7: Omitted variables bias tests for long-run teacher effects on academic outcomes

	12th grade GPA		Graduation		College bound	
	(1)	(2)	(3)	(4)	(5)	(6)
	Y	$\hat{Y}$	Y	$\hat{Y}$	Y	$\hat{Y}$
No high school	-0.0476		-0.0809		-0.0266	
	(0.00206)		(0.000654)		(0.00144)	
High school only	-0.0818		-0.0381		-0.0507	
	(0.00135)		(0.000488)		(0.000937)	
Some college	0.0220		0.0347		0.0264	
	(0.00139)		(0.000507)		(0.000963)	
BA or more	0.178		0.00226		0.118	
	(0.00111)		(0.000407)		(0.000792)	
Lag 2 math	0.106		0.00909		0.0306	
	(0.000810)		(0.000290)		(0.000566)	
Lag 2 reading	0.0455		0.00597		0.0211	
	(0.000769)		(0.000274)		(0.000536)	
Teacher effect		0.00712		0.0240		0.0132
		(0.000794)		(0.00122)		(0.000995)
Observations	3416465	3416465	4606682	4606682	3193984	3193984
R2	0.586	0.909	0.154	0.798	0.306	0.435
Original R2	.5646		.1229		.2762	
Design controls	✓	✓	✓	✓	✓	✓
Twin FE	✓		✓		✓	

Notes: This table presents tests for omitted variable bias in estimated teacher effects on long-run outcomes. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. The odd columns regress the outcome listed in the sub-header on the excluded covariates, teacher dummies, and the full set of design controls described in Section 3.1. The even columns regress predicted outcomes based on the excluded covariates on estimated teacher effects $\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$. Education variables refer to students' reported parental education, with an indicator for missing parental education data serving as the omitted category. Lag 2 math and reading refer to twice-lagged standardized test scores, with indicators for missing twice-lag scores included but not reported. Twin-effects include fixed effects for all twin pairs and an indicator for non-twin interacted with year. Results change little when regressing $\hat{Y}_{it}$ on $\hat{\alpha}_{it}$ in the sample of twins only. Original R^2 refers the R^2 of the regression without excluded covariates used to estimate teacher effects. Standard errors clustered at the student level are reported in parentheses.

Table A.8: Instrumental variables bias tests for short-run teacher effects

	Test scores		Behaviors		Study skills	
	(1)	(2)	(3)	(4)	(5)	(6)
	Schl-grd	Schl	Schl-grd	Schl	Schl-grd	Schl
$\hat{\alpha}_j$	1.002	1.052	1.255	1.681	1.083	1.126
	(0.0135)	(0.0175)	(0.270)	(1.101)	(0.0641)	(0.0825)
Observations	9779708	9779708	5422682	5422682	3404657	3404657
R^2	0.723	0.721	0.209	0.200	0.135	0.132
Design controls	✓	✓	✓	✓	✓	✓
School-grade FE	✓	✓	✓	✓	✓	✓
First stage F	47960	30550	364	24	2691	1709
P-value for $H_0 : \lambda = 1$	.875	.003	.346	.536	.198	.127

Notes: This table presents instrumental variable tests for bias in estimated teacher effects on short-run outcomes. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t + 1$, total days absent in year $t + 1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Design controls include the full set of covariates described in Section 3.1 and using all available years for each outcome. The reported coefficient on $\hat{\alpha}_{it}$ is estimated via 2SLS using a teacher switching instrument defined at the school-grade (odd columns) or school-level (even columns). The instrument is the product of an indicator for new teacher entry into student i's school-grade or school at time t times the mean of $\hat{\alpha}_j$ for all entering teachers estimated in all other school-grades or schools. Only entries where at least one new teacher's effects are estimable in other schools or school grades are included in the instrument. Means are weighted by number of students assigned at time t. All regressions include an indicator for any teacher entry. Standard errors clustered at the student level are reported in parentheses.

Table A.9: Instrumental variables bias tests for teacher-effects on criminal arrest

	Outcome: Y			Outcome: $\hat{Y}_{excluded}$		
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\alpha}_j$	1.090	1.456	1.390			
	(0.0824)	(0.386)	(0.374)			
Z_{it}				0.000191	0.000724	0.00171
				(0.000792)	(0.000797)	(0.000816)
Observations	4159500	4159500	4159500	9779708	9779708	9779708
R^2	0.0916	0.0754	0.0742	0.668	0.673	0.682
Design controls	✓	✓	✓	✓	✓	✓
School-grade FE		✓	✓		✓	✓
Dist-grade-year FE			✓			✓
First stage F	4825	1101	1207			
P-value for $H_0 : \lambda = 1$	.276	.238	.297			

Notes: This table presents instrumental variable tests for bias in estimated teacher effects on future criminal arrests. Criminal arrest refers to any non-criminal interaction with the justice system between ages 16 and 21. Design controls include the full set of covariates described in Section 3.1 and using all available years for each outcome. The reported coefficient on $\hat{\alpha}_{it}$ in columns 1-3 is estimated via 2SLS using a teacher switching instrument defined at the school-grade level. The instrument is the product of an indicator for new teacher entry into student i's school-grade or school at time t times the mean of $\hat{\alpha}_j$ for all entering teachers estimated in all other school-grades. Only entries where at least one new teacher's effects are estimable in other schools or school grades are included in the instrument. Means are weighted by number of students assigned at time t. Columns 4-6 regress the instrument on predicted outcomes using parental education and twice-lagged test scores. All regressions include an indicator for any teacher entry. Standard errors clustered at the student level are reported in parentheses.

Table A.10: Implied regression of long-run effects on short-run effects using only teachers who move across schools

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	0.009 (0.022)	-0.011 (0.018)	-0.002 (0.014)	0.006 (0.013)	-0.054 (0.049)	-0.005 (0.014)	0.025 (0.034)
Behaviors	-0.051 (0.073)	-0.033 (0.066)	-0.043 (0.048)	-0.024 (0.046)	0.277 (0.189)	0.000 (0.056)	0.103 (0.103)
Study skills	-0.014 (0.023)	0.021 (0.020)	0.000 (0.016)	-0.018 (0.014)	0.089 (0.058)	0.027 (0.016)	0.035 (0.039)
$sd(\alpha_j^y)$	0.023 (0.000)	0.019 (0.000)	0.009 (0.000)	0.014 (0.000)	0.072 (0.001)	0.012 (0.000)	0.033 (0.001)
R^2	0.026	0.045	0.108	0.038	0.084	0.095	0.077

Notes: This table presents the coefficients from a regression of long-run outcomes on short-run teacher effects implied by variance-covariance matrix of short- and long-run teachers effects. The estimates are based variance-covariance estimated from leaving-one-out teacher-school pairs rather than the leave-one-out teacher-year estimates reported in Table 4. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. The final two rows report estimate standard deviations of teacher effects on the long-run outcome and the R^2 from the regression. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome.

Table A.11: Implied regressions using within-school variation

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	-0.007 (0.006)	-0.006 (0.006)	-0.007 (0.004)	-0.003 (0.004)	0.089 (0.016)	0.009 (0.005)	0.046 (0.012)
Behaviors	-0.069 (0.030)	-0.048 (0.027)	-0.026 (0.020)	-0.018 (0.017)	0.011 (0.068)	0.018 (0.019)	0.009 (0.046)
Study skills	0.001 (0.007)	0.006 (0.007)	0.002 (0.005)	-0.005 (0.005)	0.013 (0.020)	-0.004 (0.006)	-0.001 (0.014)
$sd(\alpha_j^y)$	0.007 (0.000)	0.003 (0.000)	0.005 (0.000)	0.006 (0.000)	0.050 (0.001)	0.012 (0.000)	0.010 (0.000)
R^2	0.306	0.646	0.104	0.056	0.049	0.014	0.276

Notes: This table presents the coefficients from a regression of long-run outcomes on short-run teacher effects implied by the variance-covariance matrix of short- and long-run teachers effects using only *within* school variation in teacher effects. Estimates are based on the basline model in which teacher effects are constant across schools and time. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. The final two rows in each panel report estimate standard deviations of teacher effects on the long-run outcome and the R^2 from the regression. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome.

Table A.12: Summary statistics of four different sub-groups comparisons across race, sex, socioeconomic status, and predicted risk of arrest

	Full sample	Race		Sex		Econ. disadv.		Arrest risk	
	(1)	(2) White	(3) Non-White	(4) Boys	(5) Girls	(6) Yes	(7) No	(8) High	(9) Low
Demographics									
Male	0.50	0.51	0.49	1	0	0.50	0.51	0.72	0.28
Black	0.25	0	0.58	0.25	0.26	0.37	0.087	0.40	0.10
Economically disadvantaged	0.58	0.42	0.80	0.58	0.59	1	0	0.77	0.39
Limited English	0.043	0.0037	0.095	0.045	0.042	0.069	0.0075	0.017	0.070
Parents have HS education or less	0.40	0.33	0.50	0.40	0.40	0.55	0.19	0.56	0.23
Parents have some college	0.45	0.52	0.34	0.45	0.44	0.26	0.69	0.30	0.59
Parents have 4-year degree	0.22	0.28	0.14	0.22	0.22	0.064	0.43	0.096	0.35
Short-run outcomes									
Standardized reading scores	0.046	0.29	-0.28	-0.018	0.11	-0.26	0.48	-0.33	0.42
Standardized math scores	0.061	0.30	-0.25	0.063	0.059	-0.25	0.50	-0.30	0.43
Days absent	9.11	9.58	8.48	9.27	8.95	10.1	7.52	10.7	7.72
Any discipline	0.17	0.12	0.23	0.22	0.11	0.22	0.081	0.28	0.076
Any out-of-school suspension	0.080	0.046	0.12	0.11	0.047	0.11	0.026	0.14	0.026
Repeat grade	0.0088	0.0062	0.012	0.011	0.0062	0.013	0.0028	0.015	0.0025
Behavioral index	4.3e-10	-0.066	0.086	0.12	-0.12	0.20	-0.32	0.35	-0.25
Time spent on homework	0.023	0.089	-0.067	-0.015	0.061	-0.086	0.16	-0.12	0.14
Time spent reading	0.0052	0.045	-0.053	-0.13	0.14	-0.049	0.079	-0.16	0.14
Time spent watching TV	-0.0052	-0.16	0.20	0.061	-0.071	0.15	-0.18	0.15	-0.20
Study skills index	-6.4e-10	0.11	-0.17	-0.14	0.14	-0.15	0.20	-0.23	0.30
Long-run outcomes									
12th grade GPA (0-6 scale)	3.13	3.34	2.81	2.96	3.28	2.78	3.53	2.59	3.53
12th grade class rank	0.48	0.44	0.55	0.54	0.43	0.56	0.39	0.62	0.38
Graduate high school	0.91	0.92	0.90	0.90	0.93	0.87	0.97	0.86	0.96
Plans to attend 4-year college	0.46	0.46	0.45	0.41	0.50	0.35	0.60	0.32	0.56
Any CJC 16-21	0.44	0.43	0.47	0.52	0.37	0.48	0.39	0.53	0.36
Traffic infraction	0.33	0.33	0.34	0.39	0.28	0.35	0.31	0.39	0.29
Criminal arrest	0.24	0.21	0.28	0.30	0.17	0.29	0.16	0.34	0.15
Index crime arrest	0.10	0.078	0.15	0.13	0.083	0.14	0.047	0.16	0.052
Criminal conviction	0.10	0.084	0.13	0.15	0.054	0.13	0.052	0.16	0.043
Incarcerated	0.089	0.073	0.11	0.13	0.048	0.12	0.043	0.15	0.036
N student-subject-years	9779708	5536382	4243307	4892705	4887002	5673880	4035391	4889854	4889854
N teachers	39707	39366	39664	39702	39706	39683	39342	39688	39623
N students	1953547	1048427	905112	983078	970468	1085041	818161	1212875	1105939
N twin pairs	18213	10178	9494	11813	12072	12031	7921	13481	11390

Notes: This table presents summary statistics for demographic characteristics, short-run outcomes, and long-run outcomes for the full sample (Column 1), white vs. non-white (Columns 2 and 3), boys vs. girls (Columns 4 and 5), economic disadvantage (Columns 6 and 7), and students with high vs. low predicted risk of a future arrest (Columns 8 and 9). Not all outcomes are observed in all years; summary statistics reflect means and standard deviations for non-missing data only. In each analysis, we use the largest sample possible given when outcome studied. See Section 2 for additional details on data construction and outcome coverage by year.

Table A.13: Implied regressions using within-school variation when teacher effects vary by school

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	-0.009 (0.007)	-0.007 (0.006)	-0.008 (0.005)	-0.004 (0.004)	0.097 (0.017)	0.009 (0.005)	0.055 (0.012)
Behaviors	-0.076 (0.035)	-0.050 (0.028)	-0.030 (0.021)	-0.024 (0.019)	0.068 (0.078)	0.021 (0.021)	0.021 (0.050)
Study skills	0.006 (0.009)	0.010 (0.008)	0.003 (0.006)	-0.003 (0.006)	0.006 (0.023)	-0.005 (0.007)	-0.012 (0.017)
$sd(\alpha_j^y)$	0.010 (0.000)	0.006 (0.000)	0.006 (0.000)	0.007 (0.000)	0.053 (0.001)	0.012 (0.000)	0.014 (0.000)
R^2	0.242	0.252	0.097	0.069	0.057	0.017	0.198

Notes: This table presents the coefficients from a regression of long-run outcomes on short-run teacher effects implied by the average variance-covariance matrix of short- and long-run teachers effects using only *within* school variation in teacher effects. The estimation is based on Equation 12 that allows teacher effects to vary by school (i.e., α_{js}). Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. The final two rows in each panel report estimate standard deviations of teacher effects on the long-run outcome and the R^2 from the regression. Teacher effect estimators include the full set of covariates described in Section 3.1 and using all available years for each outcome.

B What does the covariance of EB posteriors estimate?

This appendix investigates whether covariances of latent teacher effects across outcomes or student groups can be estimated using covariances of Empirical Bayes (EB) estimates of individual teachers' effects. We show that covariances between EB posteriors can either under- or over-estimate covariances in latent effects depending on the data generating process (DGP). In cases calibrated to our data, the degree of bias can be large and either negative or positive depending on the outcomes considered.

We begin by examining the covariance of univariate EB posteriors, which are commonly used in the literature (Jackson, 2018; Petek and Pope, 2021; Backes et al., 2022; Bates et al., 2022). In the univariate case, it is simple to show analytically why the covariance of EB posteriors will differ from covariances of teacher effects due to either the shrinkage factors or correlated measurement error. Next, we examine the case of multivariate shrinkage that takes into account the covariance in teacher effects across outcomes as well as the covariance in sampling errors. We calibrate parameters to our setting and present simulations that illustrate the potential bias and how it changes with the per-teacher sample size.

Consider first a simple example using univariate EB posteriors. Suppose that outcome k of student i assigned to teacher j is determined by:

$$Y_{ij}^k = \alpha_j^k + \epsilon_i^k$$

The distribution of teacher effects α_j^k is assumed to follow:

$$\begin{pmatrix} \alpha_j^A \\ \alpha_j^C \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (\sigma_\alpha^A)^2 & \sigma_\alpha^{AC} \\ \sigma_\alpha^{AC} & (\sigma_\alpha^C)^2 \end{pmatrix}\right)$$

The distribution of the individual heterogeneity ϵ_i^k is given by:

$$\begin{pmatrix} \epsilon_i^A \\ \epsilon_i^C \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (\sigma_\epsilon^A)^2 & \sigma_\epsilon^{AC} \\ \sigma_\epsilon^{AC} & (\sigma_\epsilon^C)^2 \end{pmatrix}\right)$$

For simplicity, assume that each teacher is assigned n students with outcomes generated by this process. The average outcome of students assigned to teacher j is:

$$\bar{Y}_j^k = \alpha_j^k + \frac{\sum_{i=1}^n \epsilon_i^k 1(j(i) = j)}{n}$$

Due to the normal-normal structure of the model, the univariate EB posterior for teacher

j's effect on outcome k, or $E[\alpha_j|\bar{Y}_j]$, is simply $\lambda^k_{EB}\bar{Y}^k_j$, where

$$\lambda^k_{EB} = \frac{(\sigma^k_\alpha)^2}{(\sigma^k_\alpha)^2 + \frac{(\sigma^k_\epsilon)^2}{n}}$$

It follows immediately that the covariance of simple, univariate EB posterior means does not identify the covariance of latent teacher effects, since

$$Cov(\lambda^A_{EB}\bar{Y}^A_j, \lambda^C_{EB}\bar{Y}^C_j) = \lambda^A_{EB}\lambda^C_{EB}\left[Cov(\alpha^A_j, \alpha^C_j) + \frac{Cov(\epsilon^A_i, \epsilon^C_i)}{n}\right]$$

The direction of bias depends on two terms: attenuation due to $\lambda^A_{EB}\lambda^C_{EB}$, which falls between zero and one, and the covariance of student-level heterogeneity $Cov\left(\epsilon^A_i, \epsilon^C_i\right)$. When the latter is zero, the covariance of EB posteriors is attenuated toward zero and could in principle be corrected by undoing multiplication by $\lambda^A_{EB}\lambda^C_{EB}$. Since in general $Cov\left(\epsilon^A_i, \epsilon^C_i\right)$ can take any sign, the overall bias is unclear in general settings.

In practice, researchers may use multivariate EB estimators that take account of data for both outcomes simultaneously. These estimators also fail to recover unbiased estimates of covariances in latent effects. To show how, we construct an illustration based on the variance-covariance of teacher effects and student-level heterogeneity in our data. Table B.1 reports estimates of both, with the latter in brackets. The variances of the student-level heterogeneity are large. Indeed, in some cases they are bigger than that of teacher effects. Meaningful correlations in student-level heterogeneity of different signs across outcomes are also present.

Table B.1: Variance-covariance of latent teacher effects and student-level heterogeneity

	Test scores	Behaviors	Criminal arrest
Test scores	0.121 [0.1391]	0.056 [0.0614]	-0.008 [-0.0404]
Behaviors		0.125 [0.2144]	-0.202 [-0.1048]
Criminal arrest			0.027 [0.0697]

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects and classroom-level heterogeneity (i.e., $\frac{\sum_{i=1}^n \epsilon^k_i 1(j(i)=j)}{n}$) for key outcomes. The former is reported without brackets, while the latter is reported in square brackets.

We use the estimated variance-covariance of teacher effects and student-level heterogeneity to construct the implied covariance (and correlation) in EB posteriors from the normal-normal model described above for different values of n and examine how it relates to the true covariances (and correlations) of latent effects. The EB estimates use a multivariate

model that constructs posterior means as:

$$E[(\alpha_j^A, \alpha_j^B)|(\bar{Y}_j^A, \bar{Y}_j^B)] = \Sigma_{12}\Sigma_{22}^{-1}(\bar{Y}_j^A, \bar{Y}_j^B) \tag{B.1}$$

where $\Sigma_{12} = \begin{pmatrix} \sigma_\alpha^A & \sigma_\alpha^{AC} \\ 0 & \sigma_\alpha^C \end{pmatrix}$ and $\Sigma_{22} = \begin{pmatrix} Var(\bar{Y}_j^A) & Cov(\bar{Y}_j^A, \bar{Y}_j^B) \\ Cov(\bar{Y}_j^A, \bar{Y}_j^B) & Var(\bar{Y}_j^B) \end{pmatrix}$.

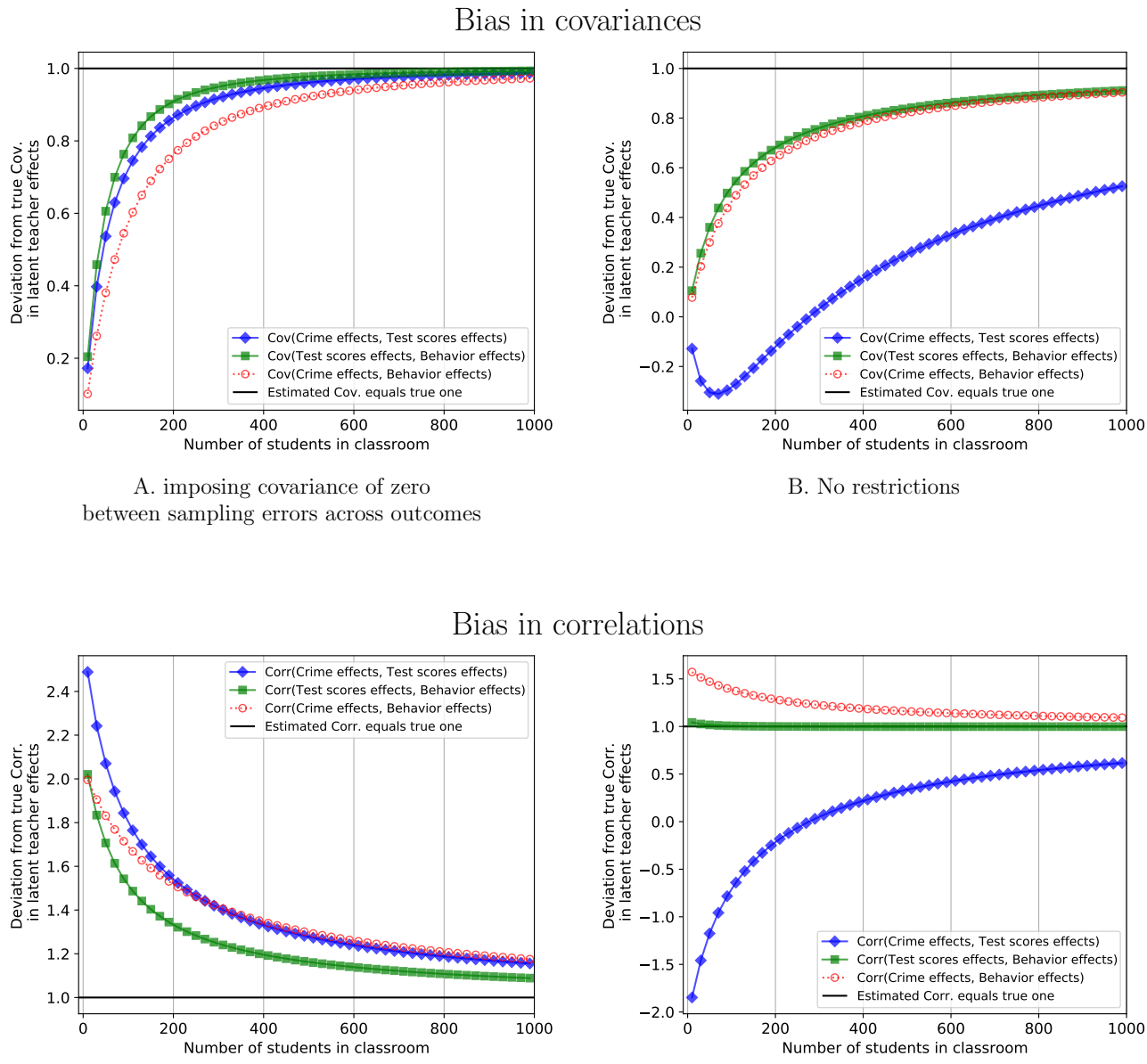
Figure B.1 reports the results. Each point in the figure shows the ratio between the covariance (or correlation) of EB posterior means, or $Cov\left(E[\alpha_j^A|(\bar{Y}_j^A, \bar{Y}_j^B)], E[\alpha_j^B|(\bar{Y}_j^A, \bar{Y}_j^B)]\right)$ and the covariance (or correlation) of latent effects, or $Cov(\alpha_j^A, \alpha_j^B)$, indicating the proportional degree of bias. The x-axis report the number of students assigned to each teacher (n).

In Panel A, we impose that there is no correlation in the student-level heterogeneity across outcomes, so that $Cov(\epsilon_i^k, \epsilon_i^{k'}) = 0$ for any two outcomes k and k'. As we would expect, the bias is decreasing with the number of students per teacher as the shrinkage factors converge to one and teacher-specific means converge to α_j^k. This pattern holds for the covariance estimates across all pairs of outcomes. However, even with 200 students per teacher, the magnitude of the bias is non-negligible.

In Panel B, we re-introduce the correlations in student-level heterogeneity reported in Table B.1. The results are consistent across the outcomes considered. The biases from correlation in student-level heterogeneity across outcomes and the shrinkage factors are large and persistent, even when there are 1,000 students per teacher. Importantly, for the covariance in test scores and criminal arrest the bias is large even with 1000 students per teacher, implying very accurate estimates of each teacher's effects.

Panels C and D report the bias of estimates of correlations rather than covariances and are analogous to Panels A and B. The results are broadly similar and show persistent and meaningful bias even with a large number of students per teacher. For the correlation between effects on behaviors and test scores, the biases from correlated student-level heterogeneity across outcomes and the shrinkage factors happen to cancel each other out. The former biases the estimate up, while the latter pulls the estimate towards zero. However, for the other correlations the biases do not cancel out. Importantly, for the correlation in test scores and criminal arrest the bias is meaningful even with 1000 students per teacher, implying very accurate estimates of each teacher's effects.

Figure B.1: Bias of the covariance of EB posterior means as an estimator of the covariance of latent teacher effects



A. imposing covariance of zero between sampling errors across outcomes

B. No restrictions

C. imposing covariance of zero between sampling errors across outcomes

D. No restrictions

Notes: This figure shows results of calculations from the data generating process described in Appendix B. The x-axis is the number of students assigned to each teacher. The y-axis compares the covariance (or correlation) in latent teacher effects across outcomes A and B (e.g., test scores and criminal arrest between ages 16 to 21) to the covariance (or correlation) in EB posterior means. Each dot plots the ratio of covariances (or correlations), $Cov(\alpha_j^A, \alpha_j^B)$ to $Cov\left(E[\alpha_j^A|(\bar{Y}_j^A, \bar{Y}_j^B)], E[\alpha_j^B|(\bar{Y}_j^A, \bar{Y}_j^B)]\right)$. Panel A imposes that there is no correlation in student-level heterogeneity, while Panel B uses estimates from our data reported in Table B.1. Panels C and D are analogous to Panels A and B but report results for correlations rather than covariances.

C Inference on variance components

Our standard errors rely on second-order U-statistic representations of the estimators in Equations 4 and 7. Specifically, the estimator for the covariance in latent effects between two outcomes (e.g., A and C) or the variance of latent effects (e.g., when $A = C$) can be written as:

$$\widehat{Cov}(a_j^A, a_j^C) = \sum_i \sum_{k \neq i} C_{ik}^{AC} Y_i^A Y_k^C$$

$$C_{ik}^{AC} = \begin{cases} \frac{J-1}{J^2} \frac{1}{|T_j^A||T_j^C| - |T_j^A \cap T_j^C|} & \text{if } j(i) = j(k) \\ \frac{-1}{|T_{j(i)}^A||T_{j(k)}^C| J^2} & \text{if } j(i) \neq j(k) \end{cases}$$

where, with a slight abuse of notation, i and k index teacher-year mean residuals, i.e., $Y_i^k = \bar{Y}_{j(i)t(i)}^k$, if outcome k is observed for teacher j in year t and zero otherwise, J is the total number of teachers, and T_j^k is the set of time periods where outcome k is observed for teacher j.

The sampling covariance between any two covariance (or variance) estimates of effects on outcomes A and B and C and D can be expressed as:

$$\begin{aligned}
&Cov\left(\widehat{Cov}(a_j^A, a_j^B) - Cov(a_j^A, a_j^B), \widehat{Cov}(a_j^C, a_j^D) - Cov(a_j^C, a_j^D)\right) \\
&= Cov\Bigg(\sum_i v_i^A \sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^B + \sum_i v_i^B \sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^A + \sum_i v_i^A \sum_{k \neq i} C_{ik}^{AB} v_k^B, \\
&\quad \sum_i v_i^C \sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^D + \sum_i v_i^D \sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^C + \sum_i v_i^D \sum_{k \neq i} C_{ik}^{CD} v_k^C \Bigg) \\
&= \sum_i \sigma_i^{AC} \left(\sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^B \right) \left(\sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^D \right) + \sum_i \sigma_i^{AD} \left(\sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^B \right) \left(\sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^C \right) \\
&+ \sum_i \sigma_i^{BC} \left(\sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^A \right) \left(\sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^D \right) + \sum_i \sigma_i^{BD} \left(\sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^A \right) \left(\sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^C \right) \\
&+ \sum_i \sigma_i^{AD} \sum_{k \neq i} C_{ik}^{AB} C_{ik}^{CD} \sigma_k^{BC} + \sum_i \sigma_i^{AC} \sum_{k \neq i} C_{ik}^{AB} C_{ik}^{CD} \sigma_k^{BD}
\end{aligned}$$

where as in the main text we define $\bar{Y}_{j(i)t(i)}^k = a_{j(i)}^k + \bar{v}_{j(i)t(i)}^k$ and define σ_i^{AB} as the covariance between $\bar{v}_{j(i)t(i)}^A$ and $\bar{v}_{j(i)t(i)}^B$.

Special cases of this expression deliver sampling variances for objects such as $Var(\hat{a}_j^A)$, which

can be written as:

$$Var(\widehat{Var}(a_j^A) - Var(a_j^A)) = 4\sum_i (\sigma_i^A)^2 \left(\sum_{k\neq i} C_{ik}^{AA} a_{j(k)}^A\right)^2 + 2\sum_i \sum_{k\neq i} (C_{ik}^{AA})^2 (\sigma_i^A)^2 (\sigma_k^A)^2$$

where $(\sigma_i^A)^2 = \sigma_i^{AA}$. Standard errors for the covariance in latent effects between two outcomes (e.g., A and C) are also a special case and can be written as:

$$\begin{aligned} Var(\widehat{Cov}(a_j^A, a_j^C) - Cov(a_j^A, a_j^C)) = &\sum_i (\sigma_i^A)^2 \left(\sum_{k\neq i} C_{ik}^{AC} a_{j(k)}^C\right)^2 + \sum_i (\sigma_i^C)^2 \left(\sum_{k\neq i} C_{ik}^{AC} a_{j(k)}^A\right)^2 \\ &+ 2\sum_i (\sigma_i^{AC})^2 \left(\sum_{k\neq i} C_{ik}^{AC} a_{j(k)}^A\right)\left(\sum_{k\neq i} C_{ik}^{AC} a_{j(k)}^C\right) \\ &+ \sum_i \sum_{k\neq i} (C_{ik}^{AC})^2 (\sigma_i^A)^2 (\sigma_k^C)^2 + \sum_i \sum_{k\neq i} (C_{ik}^{AC})^2 \sigma_i^{AC} \sigma_k^{AC} \end{aligned}$$

C.1 Plug-in estimator of standard errors

As noted in Kline, Saggio and Sølvsten (2020), plug-in estimators of these variances using and $\hat{a}_j^k$ and $\hat{\sigma}_j^{kl}$ will generically be biased since, for example, $(\hat{\lambda}_i^{ml})^2 = \left(\sum_{l\neq i} C_{ik}^{ml} \hat{a}_{j(l)}^l\right)^2$ is not an unbiased estimate of $(\lambda_i^{ml})^2 = \left(\sum_{l\neq i} C_{ik}^{ml} a_{j(l)}^l\right)^2$. It is straightforward to construct a correction for these terms, however, since for example $E[(\hat{\lambda}_i^{ml})^2] - Var(\hat{\lambda}_i^{ml}) = E[\hat{\lambda}_i^{ml}]^2$, and:

$$Var(\hat{\lambda}_i^{ml}) = (C_{jj}^{ml})^2 (\sigma_i^l)^2 \frac{(|T_{j(i)}^l| - 1)^2}{|T_{j(i)}^l|} + \sum_{s\neq j(i)} (C_{js}^{ml})^2 (\sigma_s^k)^2 |T_{ls}^l|$$

We can express the bias correction for the product of λ_i^{ml} and λ_i^{gh} analogously. We use these corrections to construct unbiased estimates of $(\lambda_i^{ml})^2$ and $\lambda_i^{ml}\lambda_i^{gh}$. We use the same unbiased estimates of $(\sigma_i^k)^2$ and σ_i^{kl} as in the main text to form our plug-in estimators of sampling variances and covariances. The remaining bias in the resulting plug-in estimate of sampling variances stems from terms such as $(\hat{\sigma}_i^k)^4$. Though it is possible to construct unbiased estimates of these objects using split sample techniques, we do not do so. As discussed in Kline, Saggio and Sølvsten (2020), using the biased plug-in versions of these terms results in conservative inference but avoids the need to subset to teachers with sufficient observations to construct split sample estimates of these terms.

C.2 Standard errors for functions of variance-covariance estimates

Wherever possible, we use the delta method to construct standard errors for functions of multiple variance-covariance components, such as a correlation coefficient. In some cases, however, we use a parametric bootstrap assuming that variance-covariance estimates are normally distributed around the point estimates, with sampling variance-covariance structure given by estimated sampling variance-covariances. Doing so provides a convenient way to generate standard errors for more complicated objects, such as multivariate regression coefficients.

D Policy simulation details

This appendix includes technical details for the implementation of the simulations discussed in Section 6. These simulations examine the implications for a given long-run outcome of replacing the bottom five percent of teachers, according to a given measure of quality, with an average teacher for exposed students. In Section 6, we discuss ranking teachers based on three different options: (i) an index based on teachers' true direct effect on long-run outcomes, (ii) an index using teachers' true effects on short-run outcomes, and (iii) an index using Empirical Bayes estimates of teacher effects on short-run outcomes. In all cases, we assume that all short- and long-run teacher effects are jointly normally distributed, allowing us to characterize the full distribution of teacher quality using our variance estimates.

D.1 Index using true teacher effects on long-run outcomes

In this case, the calculations are straightforward. We are interested in the impact of replacing the bottom five percent of teachers according to the quality index in Equation 13 with the average teacher. Thus, when estimating the effect of such a policy on teachers' effect on college attendance, for example, the estimand of interest is:

$$E[\mu^A] - E[\mu^A | \omega\mu^C + (1-\omega)\mu^A < q_{0.05}^{\text{Ideal long-run}}]$$

where $q_{0.05}^{\text{Ideal long-run}}$ is the fifth percentile of the distribution of $\mu^C + (1-\omega)\mu^A$. The calculation is straightforward given the properties of a bivariate normal distribution and the variance-covariance matrix of $(\mu^A, \omega\mu^C + (1-\omega)\mu^A)$.

D.2 Index using true teacher effects on short-run outcomes

In this case, the calculations are also straightforward. We are interested in the impact of replacing the bottom five percent of teachers according to the quality index in Equation 14 with average teachers. Thus, when estimating the effect of such a policy on teachers' effect on college attendance, for example, the estimand of interest is:

$$E[\mu^A] - E[\mu^A | \omega_1\mu^T + \omega_2\mu^B + (1-\omega_1-\omega_2)\mu^S < q_{0.05}^{\text{Ideal short-run}}]$$

where $q_{0.05}^{\text{Ideal short-run}}$ is the fifth percentile of the distribution of $\omega_1\mu^T + \omega_2\mu^B + (1-\omega_1-\omega_2)\mu^S$. The calculation is straightforward given the properties of a bivariate normal distribution and the variance-covariance matrix of $(\mu^A, \omega_1\mu^T + \omega_2\mu^B + (1-\omega_1-\omega_2)\mu^S)$.

D.3 Index using Empirical Bayes estimates of effects on short-run outcomes

In this case, the calculations require a few steps. Recall that we are interested in the impact of replacing the bottom five percent of teachers with average teachers according to an Empirical Bayes estimate of the quality index in Equation 14.

The policy maker observes the performance of the students of teacher j along multiple dimensions: $(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)$. Thus, the first step is forming an Empirical Bayes estimate of the teacher quality index. We assume that all random variables are normally distributed and that teacher effect estimates are the sum of true teacher effects and independent, identically distributed noise. The Empirical Bayes estimate is:

$$\text{Index}_j^{\text{EB short-run}} = E[\underbrace{\omega_1\mu^T + \omega_2\mu^B + (1-\omega_1-\omega_2)\mu^S}_{=\text{Index}_j^{\text{Ideal short-run}}}|\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S] \tag{D.1}$$

$$= E[\text{Index}_j^{\text{Ideal short-run}}] + \boldsymbol{b}_I' \left[(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S) - E[(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)]\right] \tag{D.2}$$

where $\boldsymbol{b}_I$ is the linear projection of $\text{Index}_j^{\text{EB short-run}}$ on $(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)$, i.e., $\Sigma_{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}}^{-1}\Sigma_{\hat{\boldsymbol{\mu}}\text{Index}_j^{\text{EB short-run}}}$ with $\hat{\boldsymbol{\mu}} = (\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)$. The last equality follows from the properties of the multivariate normal distribution. Note, that as $\text{Index}_j^{\text{EB short-run}}$ is a linear combination of normally distributed variables, then it is also normally distributed.

The second step is to predict the effect of conducting a policy that replaces all teachers with $\text{Index}_j^{\text{EB short-run}}$ that is below the 0.05 percentile with the average teacher. Since $\text{Index}_j^{\text{EB short-run}}$ is normally distributed, calculating its fifth percentile is straightforward.

To formulate our best predictor of the impact of the policy on an outcome of interest Y, we use also teachers' observed performance. We construct our estimator in two steps. First, we calculate the Empirical Bayes estimate of Y given $(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)$:

$$\hat{Y}^{\text{EB short-run}} = E[Y|\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S] \tag{D.3}$$

$$= E[Y] + \boldsymbol{b}_Y' \left[(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S) - E[(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)]\right] \tag{D.4}$$

where $\boldsymbol{b}_Y$ is the linear projection of Y onto $(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)$. The second step is to calculate the

predicted change in Y due to the replacement policy:

$$\underbrace{E[\hat{Y}^{\text{EB short-run}}]}_{=E[Y]} - E[\hat{Y}^{\text{EB short-run}} | \text{Index}_j^{\text{EB short-run}} < q_{0.05}^{\text{EB of short-run}}] \tag{D.5}$$

$$= \frac{Cov(\hat{Y}^{\text{EB short-run}}, \text{Index}_j^{\text{EB short-run}})}{Var\left(\text{Index}_j^{\text{EB short-run}}\right)} E[\text{Index}_j^{\text{EB short-run}} | \text{Index}_j^{\text{EB short-run}} < q_{0.05}^{\text{EB of short-run}}]$$

$$= \frac{Cov(\hat{Y}^{\text{EB short-run}}, \text{Index}_j^{\text{EB short-run}})}{Var\left(\text{Index}_j^{\text{EB short-run}}\right)} \sigma_{\text{Index}_j^{\text{EB short-run}}} \frac{\phi\left(\frac{q_{0.05}^{\text{EB of short-run}} - E[\text{Index}_j^{\text{EB short-run}}]}{\sigma_{\text{Index}_j^{\text{EB short-run}}}}\right)}{\Phi\left(\frac{q_{0.05}^{\text{EB of short-run}} - E[\text{Index}_j^{\text{EB short-run}}]}{\sigma_{\text{Index}_j^{\text{EB short-run}}}}\right)}$$

and note that:

$$Cov(\hat{Y}^{\text{EB short-run}}, \text{Index}_j^{\text{EB short-run}}) = Cov(\boldsymbol{b}_I' \hat{\boldsymbol{\mu}}, \boldsymbol{b}_Y' \hat{\boldsymbol{\mu}})$$

Implementing this homoscedastic EB-version of retention policies requires only estimating the variance-covariance of $(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)$, which can be directly estimated given individual teacher effect estimates, and the previously estimated variance-covariances of teacher effects on short-run outcomes.