The Effects of Teacher Quality on Adult Criminal Justice Contact

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

December 2021

Abstract

This paper investigates the impact of teacher quality on future criminal justice contact and long-run academic outcomes. Using a unique data set linking the universe of North Carolina public school records to administrative arrest and court records, we measure the impacts of teachers on short-run cognitive and non-cognitive outcomes, including test scores, school discipline, and study skills. We find that teachers' test score impacts are orthogonal to their impacts on students' likelihood of future arrest, but are correlated with academic outcomes such as college attendance. Teachers who reduce suspensions and improve attendance both meaningfully reduce future arrests and improve long-run academic outcomes. Teacher quality measures based on tests scores alone therefore miss an important component of the social value of good teachers. Because effects on short-run outcomes have different correlations with each long-run outcome, quality measures that emphasize test scores vs. other dimensions also trade off emphasis on various long-run outcomes. Regardless of the outcome, however, a pure test-score based measure is dominated by measures that incorporate effects on behaviors.

[*] Evan K. Rose: Saieh Family Research Fellow, University of Chicago; ekrose@uchicago.edu. Jonathan Schellenberg, Associate at Cornerstone Research; jschellenberg@econ.berkeley.edu. Yotam Shem-Tov, Assistant Professor Department of Economics, 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, 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, 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). The accumulation of non-cognitive and socio-emotional "soft skills" may lie at the heart of the return to education for crime (Heckman and Rubinstein, 2001; Heckman, Stixrud and Urzua, 2006; , 2010; Heckman and Kautz, 2012; Heckman, Pinto and Savelyev, 2013; Jackson et al., 2020). Consistent with this idea, policy changes and interventions that meaningfully decrease CJC typically target a child's decision making process rather than improve academic achievements (e.g., Heller et al., 2016).

Related research has focused on a central component of educational quality: the quality of educators. Measured by their influence on students' test scores, teacher quality 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. Chetty, Friedman and Rockoff (2014*b*), for example, find that replacing the five percent of teachers worst at increasing test scores would boost students' lifetime earnings by about $250,000 per exposed classroom. 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, such as criminal activity. Teachers' effects likely also vary across students. Measuring what makes a teacher "good" may therefore depend both on the long-run outcome considered and the characteristics of the student. Performance-pay and retention policies (e.g., Neal, 2011) that rely on overall effects for particular short-run outcomes like test scores may neglect teachers' impacts on other, high-stakes long-run outcomes or on particular types of students.

This paper begins by estimating teachers' impacts on standardized test scores in math and reading, a set of behavioral proxies for non-cognitive skills such as attendance and disciplinary infractions, and measures of study skills. We then ask whether teachers who improve each outcome affect their students' likelihood of arrest, conviction, and incarceration as young adults. We contrast these results with teachers' impacts on long-run academic outcomes, including graduation and plans for four-year college attendance. We first consider teachers' impacts on the average student, then allow for heterogeneous effects by students' demographic and academic characteristics and school. Our empirical approach relies on direct estimators of the variance-covariance structure of latent teacher effects and is bolstered by a battery of specification tests using excluded covariates and teacher switches across schools and grades. We conclude by showing that quality measures that emphasize different short-run impacts trade off effects along long-run dimensions, but policies based solely on test scores are dominated by those that incorporate other impacts regardless of the outcome considered.

The analysis relies on a unique merge of administrative criminal justice and education data sets in North Carolina that include more than 1.7 million students and over 33,000 teachers. The education records cover all students in N.C. public schools in grades 3 to 12 from the mid-1990s to 2013 and include rich data on students and their outcomes. We make use of three particularly important variables—students' parental education, twice-lagged test scores, and indicators for twins—to test teacher effect estimates for bias. 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 is high quality.

We estimate substantial variation in teachers' impacts on short-run outcomes. The standard deviation of teacher effects on math test scores, for example, is 0.14. Recent research in North Carolina, Los Angeles, and New York has found similar figures. Teachers also strongly influence students' homework and reading time outside the classroom, as well as the likelihood students are suspended, come to class, and are forced to repeat a grade. The standard deviation of teachers' effects on their students' chances of out of school suspensions in the next year, for example, is 0.026, or 34 percent of the mean suspension rate. 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.

We find that teachers who boost test scores do not meaningfully decrease the chances their students encounter the justice system 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 statistical 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*).

On the other hand, teachers' impacts on behavioral outcomes are closely connected to their impacts on students' future CJC. A one standard deviation increase in teacher quality measured by their 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. Compared to the mean frequency of the outcome, effects are largest on more serious criminal events: Index crimes and incarceration respond the most, while traffic tickets respond the least.[1] If teacher effects on short-run behaviors are evidence of influence over non-cognitive skills and traits such as conscientiousness, perseverance, and sociability (Heckman, Stixrud and Urzua, 2006; Lleras, 2008; Bertrand and Pan, 2013), the evidence supports that teachers' effects on CJC flow through these channels.[2]

While the estimates rely on observational variation in teacher assignments and are thus in principle vulnerable to selection bias, the highly detailed administrative records make it possible to explicitly control for a wide range of confounds. Multiple tests demonstrate that teacher effects on both short- and long-run outcomes 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 model controls. Using teachers switches across schools and school-grades to instrument for students' exposure to particular teachers, we cannot reject that estimates are unbiased. That is, teachers' effects on outcomes when they move align closely with what one would predict based on their effects estimated in other schools and grades.

Moreover, estimates of the long-run effects of teachers' impacts on test scores and behaviors show similar patterns to the impacts of boosting test scores or changing behaviors measured within twins. While test scores are strongly related to CJC conditional on our standard controls, within twins the estimated effect of increasing test scores on CJC is significantly attenuated, suggesting much smaller gains to increasing tests scores than observational relationships imply. The effects of improving behaviors on future CJC within twins are similar to the effects of assignment to teachers with strong behavioral effects. These results suggest that the skills required to excel in these short-run outcomes have different bearing on future CJC regardless of how they are acquired.

If our identifying assumptions hold for long-run outcomes, it is also possible to examine the total variance of teachers' impacts on their students' future CJC. We find a standard

[1] Index crimes are the eight crimes the FBI combines to produce its annual crime index. These offenses include homicide, forcible rape, robbery, burglary, aggravated assault, larceny, motor vehicle theft, and arson.

[2] 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).

deviation of teacher effects on future arrests of 2.8 p.p. (11.7 percent of the sample mean) and on incarceration of 2.2 p.p. (24.4 percent). Notably, teacher effects on all short-run outcomes explain no more than 4 percent of their effects on criminal arrests. Teachers therefore impact students in myriad ways not well measured by *any* short-run outcomes. While the models underlying these estimates differ than those for short-run outcomes, which enter as lags, we show that under weaker assumptions our estimates provide lower bounds for the variance of direct effects.

Even if teacher effect estimates are unbiased, good teachers may be concentrated in particular schools. Our variance estimates reflect both within-school teacher comparisons and cross-school teacher sorting. A decomposition shows that 60 to 80 percent of study skill and test score effect variances occurs within schools, while only about 20 percent for behaviors and long-run outcomes. Nevertheless, variance-covariance estimates based on purely within-school comparisons yield similar relationships between short- and long-run effects. Thus while the *total* variance of teacher effects depends on whether one compares teachers working in the same or different schools, the core conclusion that behavioral effects are related to CJC while test score effects are not remains unchanged. In addition, allowing teacher effects to vary by school changes results little, an important finding given the centrality of school-level policies for disciplinary outcomes (Bacher-Hicks, Billings and Deming, 2019; Sorensen, Bushway and Gifford, 2019).

It is also possible that teachers good at reaching particular types of students both increase their academic achievement and reduce their risk of arrest. Some work, for example, finds that same-race teachers improve students' short- and long-run outcomes (Dee, 2005; Gershenson et al., 2018). Interestingly, we find that teacher impacts on short-run outcomes are tightly correlated across multiple dimensions of heterogeneity, including sex, race, socioeconomic status, and predicted arrest risk based on all covariates in the data. The correlation in teachers' test score effects for boys and girls is 0.96, for example. Teacher effects on behavioral outcomes are less correlated across groups, but they remain tightly linked.

As a result, assignment to a teacher good at boosting test scores or improving behaviors for demographically similar students is comparable to the average effect. Hence there is limited scope to improve long-run outcomes by re-sorting students across teachers on the basis of their short-run effects. Examining effects on long-run outcomes directly, however, we find much weaker correlations. The correlation of teacher effects on white and non-white students criminal arrests is roughly 0.5, for example. Teachers' impacts on students not measured by short-run outcomes therefore exhibit considerably more heterogeneity than their impacts on test scores, behaviors, and study skills. Understanding the sources and short-run predictors of these impacts is an important area for future research.

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. While putting emphasis on different short-run outcomes trades off effects on long-run academic and CJC outcomes, any solely test score-based rule is Pareto dominated by a policy that evaluates teachers using their effects on behaviors as well.

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 estimate teacher effects. Section 4 presents and validates the results. Section 5 examines the importance of schools. In Section 6, we examine heterogeneity in teacher effects based on student characteristics. In Section 7, we simulate the effects of various retention policies. Section 8 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 1997 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.[3]

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 principle 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 principle 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 $t+1$).[4] 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

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

[4] 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.

these proxies for the relevant latent factors, an interesting direction for future research.

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 principle 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.[5]

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.[6]

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

[6] 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 first name and date of birth.[7] 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.[8]

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).[9]

2.5 Summary statistics

Summary statistics for the final analysis sample are presented in Table 1. The sample includes 7,422,917 student-subject-year observations for 1,776,759 students with 33,880 different teachers. Roughly 25 percent of the sample are Black—close to the N.C. population average, 59 percent are economically disadvantaged and receive a subsidized lunch, 22 per-

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

[8] 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.

[9] The analysis sample described below also implements an additional restriction based on a pre-testing procedure detailed in Section 3.1. This restriction removes roughly 20 percent of the sample.

cent have a parent with a four-year college degree, while 40 percent have a parent with a high-school education or less.

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 the average math and reading test scores are slightly higher (0.049 and 0.038), primarily due to the exclusion of students without a lag score. 16 percent of the students have some disciplinary infraction in an average year and 7.7 percent have an out-of-school suspension. The average 12th grade GPA in the sample is 3.12 (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 (45 percent) as students whose parents have exposure to any college (44 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 before the age of 21: 11 percent are arrested for an index crime, ten percent are convicted of a crime, and nine percent are incarcerated (including both jail and prison).

Table 1 also reports summary statistics for the sample of children who have a criminal arrest between ages 16 and 21. These students are more likely to be eligible for free or reduced-price lunch, 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. 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. To begin, we assume that there is no heterogeneity in teacher effects, allowing us to write $\epsilon_{ijt} = \epsilon_{it}$.

We define "observational" teacher effects as the population projection version of Equation 1:

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

Equation 2 is a projection defined by the population requirement that $E[D_{ijt}u_{it}] = 0$. 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:

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

Conditional independence does not require random assignment of students to teachers. Instead, teacher assignments must be uncorrelated with unobserved factors that influence outcomes conditional on observables X_{it}. This 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. It 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. We discuss several tests for violations of this assumption due to unobserved sorting below.

To support Assumption 1, 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, free lunch eligibility, 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.

To further support Assumption 1, we also implement a pre-testing procedure that identifies school-grades where students' lagged teacher assignments predict current teacher assignments, evidence of potential tracking that may be more difficult to account for with controls. This test is illustrated in Appendix Table A.2. Our main estimates drop the roughly 20 percent of school-grades where this test fails. Importantly, however, tracking of students to particular teachers across grades is not necessarily inconsistent with Assumption 1. Indeed, we show later that results change little when including all the data.

Despite the detailed controls, conditional independence may nevertheless be 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.[10] Observational effects are therefore unbiased linear predictors of causal effects and teachers with the same α_j all have expected μ_j equal to α_j. Naturally, conditional independence implies forecast unbiased effects. When only the latter holds, however, any individual teacher's causal effects may be estimated with bias. Despite this, many parameters of interest can still be estimated or bounded, as we discuss below.

The preceding discussion considered a generic outcome Y_{it}. In what follows, we consider

[10] In practice, α_j is estimated with error and $\hat{\alpha}_j$ is not a forecast unbiased predictor of μ_j even if Assumption 2 holds. We detail our solution to this issue, typically overcome using empirical Bayes techniques, below.

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 population 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 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). In doing so, we forgo development of empirical Bayes estimates of teacher-specific effects, which the prior literature typically uses in regression analyses to study impacts of teacher quality on long-run outcomes. Our approach allows us to non-parametrically characterize the distribution of teacher effects without homoscedasticity assumptions, the use of an intermediate shrinkage step, or specifying a complete statistical model.[11] We return to the question of how to obtain good predictions of any given teacher's effects in the final section of the paper.

[11] As we show below, however, using a multi-step approach as in Chetty, Friedman and Rockoff (2014*a*) ultimately produces similar conclusions.

We define the variance of teacher effects (for a generic outcome) as:

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

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 $E[\bar{v}_{jt}] = 0$.[12] We assume $\bar{v}_{jt}$ is uncorrelated across years, which requires that any cohort or school-level shocks are independent over t.[13]

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

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

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

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).[14]

[12] $\hat{\Gamma}$ 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.

[13] 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 millions of observations in the sample, any estimation error in Γ is likely to be very small, mitigating this concern.

[14] 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

$\widehat{Var}(\alpha_j)$ 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. For test score and CJC effects—α_j^A and α_j^C—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 direction. Our covariance estimator is constructed assuming that Assumption 3 holds across outcomes and excludes products of residuals from the same year, much like $\widehat{Var}(\alpha_j)$:

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

As with the variance estimator, $\widehat{Cov}(\alpha_j^A,\alpha_j^C)$ is numerically equivalent to estimating the covariance of OLS estimates across outcomes and subtracting a correction for within-teacher

the expectation using the median value in 200 Monte Carlo simulations. Our estimator uses all possible pairs of years within a teacher and therefore does not require an approximation using simulations of random matches of pairs of years t and t' within each 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 $\widehat{Cov}_j^{A,C}=\frac{1}{T_j-1}\sum_{t=1}^{T_j}\left(Y_{jt}^A-\bar{Y}_j^A\right)\left(Y_{jt}^C-\bar{Y}_j^C\right)$.[15]

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 is the same as 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.[16] This expression also makes the direction of any potential bias clear. If, for example, teachers who have the largest causal effects on test scores tend to be assigned students least likely to be arrested, the observational correlation will be more negative than the causal correlation.

Although we primarily interpret our observational correlations as causal correlations, we

[15] Note that the estimators in Equations 7 and 8 will no longer be numerically equivalent if there are missing values of one of the outcomes in some years.

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

discuss the signs of any potential biases while describing the results.

4 The causal effects of teachers

This section analyzes teacher effects on short- and long-run outcomes and their correlation structure. We then compare our estimates to effects of short-run outcomes measured in alternative designs. Finally, we conduct multiple tests of Assumptions 1 and 2 and demonstrate robustness to alternative specifications.

4.1 Short-run effects

Table 2 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. Standard errors from a weighted bootstrap procedure are included in parentheses and described in more detail in Appendix B. The estimates combine all grades four to eight.

The estimated standard deviation of teacher effects on test scores—combining homeroom, math, and reading teachers—is 0.123. 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 12.3 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.[17]

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 0.681, 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 2 shows wide variation in teacher effects on study skills. The estimated standard deviation is 0.19. Unsurprisingly, study skills effects are correlated with effects on test scores (0.314), suggesting teachers whose students complete more homework and substitute from watching television to reading also tend to see increases in test scores.

[17] Appendix Figure A.1 shows these estimates are comparable in magnitude to figures from other studies in the literature.

The fifth column shows that the estimated standard deviation of teacher effects on behaviors is 0.129. 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 12.9 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 0.067. Similarly, behaviors and study skills effects are only weakly related with a correlation of 0.044.

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 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.[18]

Appendix 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 0.756, respectively. Though somewhat difficult to compare due to differences in normalization and definitions, these estimates are roughly comparable to those in recent literature.

Dynamics and fade-out. Consistent with results in prior studies (e.g., Jacob, Lefgren and Sims, 2010; Chetty, Friedman and Rockoff, 2014*a*; Petek and Pope, 2021), we find that teacher effects on test scores fade out relatively quickly over time. Appendix Figure A.2 presents estimates of the correlation of teacher effects on outcomes in year t with effects on outcomes in subsequent years. If teachers generated permanent changes in test scores, we would expect high correlations in subsequent years. Instead, the correlation decreases

[18] 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 an opposite pattern: a strong correlation between teacher effects on learning skills and a 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. As we discuss more below, estimates of the correlations between teacher effects on short- and long-run outcomes indicate that teachers who improve study skills do not improve long-run outcomes.

sharply. The correlation in teacher effects on test scores in the current year and three years later, for example, is less than 0.3. In contrast, the correlation between teacher effects in the current year and next year is 0.9 for behaviors. However, over time effects on behavior also fade out, although less quickly, and after three years, the correlation is over 0.4.

The relatively sharp fade-out of short-run effects presents a puzzle for teacher effects on long-run outcomes documented in this and previous studies (e.g., Chetty, Friedman and Rockoff, 2014*a*). If teachers have permanent effects on students' skills that increase earnings or college attendance and decrease CJC, they are not reflected in permanent changes in our proxies for them. It is not surprising therefore that short-run measures of teacher quality have a low overall predictive ability for effects on long-run outcomes, as we show in the following subsection. While teacher effects on behaviors seem to be more persistent, they are also uncorrelated with teachers' effects on test scores. Thus if the long-run impacts of assignment to high test-score effect teachers flows through permanent changes in non-cognitive skills, these skills are not captured by our behavioral measures.

4.2 Connecting short- and long-run effects

Table 3 reports estimates of the correlation in teacher effects on short- and long-run outcomes, or $\frac{Cov(\alpha_j^A,\alpha_j^C)}{\sqrt{Var(\alpha_j^A)\cdot Var(\alpha_j^C)}}$ for outcomes A and C. The first row of Panel (a) shows that test score effects are weakly correlated with effects on any CJC, criminal arrests, arrests for an index crime, and incarceration. The estimated correlation in teacher effects on tests scores and criminal arrests, for example, is -0.002 and is not statistically significant. Although this estimate implies that teachers who increase test scores decrease arrests, the correlation is sufficiently small that we cannot reject that teacher impacts on CJC and academic achievement are orthogonal. Panel (b), however, shows that test score effects have a stronger correlation with 12th grade GPA, high school graduation, and college attendance.

The second row of Panel (a) shows that teacher impacts on behaviors are much more tightly connected to their impacts on CJC. The estimated correlation between teacher effects on the summary index of behaviors and criminal arrest is -0.214. Teachers who promote the development of "non-cognitive" skills that help students stay in school and out of trouble are therefore also very likely to help their students avoid arrest, conviction, and incarceration as young adults. Interestingly, teachers' behavioral effects are also positively correlated with effects on 12th grade GPA, high school graduation, and college attendance (Panel (b)), indicating that these skills are also important for long-run academic outcomes.

Figure 1 translates these estimates into the implied effects of exposing a student to a teacher with one standard deviation higher effects on the short-run outcome. We do so by rescaling estimated covariances to recover the implied coefficient from a regression of long-run on short-run effects, then multiplying by an estimate of the standard deviation of the latter.[19] The signs of effects are normalized so that the increase 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 of a one standard deviation shift 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* (i.e., unobserved) 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 issue of estimating teacher-specific short-run impacts 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 0.5 percent. For the purposes of recruiting, retaining, or rewarding teachers likely to help their students avoid criminal careers, therefore, test score value-added is not a particularly useful predictive measure.

This null result is not a feature of our data or estimation strategy. Panel (b) shows 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 0.65 p.p., similar to the estimated impact on actual college attendance in Chetty, Friedman and Rockoff (2014*b*) of 0.86 p.p.). Appendix Table A.3 shows that we find similar patterns when using conventional "shrinkage" methods by regressing students' future CJC on assigned teachers' value-added estimated in a multi-step procedure that allows for drift. The effect of assignment to a teacher with one standard deviation higher test score value-added on future arrests, for example, is 0.09 p.p.

By contrast, 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 2.5 percent decrease in the likelihood of a criminal arrest and a 3.8 percent reduction

[19] That is, the regression coefficient for short-run outcome A and long-run outcome C is given by $\frac{Cov(\alpha_j^C,\alpha_j^A)}{Var(\alpha_j^A)}$. Using the estimators in Equations 4 and 7, it is simple to obtain an estimate of the regression coefficient.

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. As with test scores, Appendix Table A.4 shows results change little when using alternative estimators that account for drift.[20] Teachers' impacts on behaviors also affect long-run academic outcomes, with similar effects as test score impacts on 12th grade GPA, for example.[21]

Teachers' effects on study skills, on the other hand, show inconsistent patterns of effects on long-run CJC and academic outcomes, seemingly increasing arrests but decreasing incarceration, for example. We show later that these estimates are particularly sensitive to the variation used. When solely exploiting comparisons across students and teachers in the same school, for example, study skills impacts show no significant effects on any long-run outcome. We therefore find little evidence of a robust pattern of effects for study skills.

4.3 Comparison to effects within twins

To gauge the magnitude of these estimated effects, it is useful to compare estimates to the observational relationship between tests scores, behaviors and CJC. Figure 2 presents these relationships. Panel (a) shows that grade four test scores are strongly negatively correlated with future criminal arrests. Without any additional controls, increasing test scores by one standard deviation is associated with a 5.74 p.p. decrease in future arrests, an effect orders of magnitude times larger than the estimated effect of assignment to a one standard deviation better test score effect teacher.

A large component of the observational relationship between test scores and CJC, however, likely reflects selection on unobserved confounds. Panel (b) estimates the same relationship in the sample of twins with family fixed effects included. The relationship between test scores and CJC is attenuated significantly, falling to -1.3 p.p. effects, suggesting much smaller gains to increasing tests scores than observational relationships imply.

The similarly small effects on future CJC of test scores within twins and of assignment to

[20] 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.

[21] Appendix Figure A.3 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.

high test score effect teachers suggests that the skills required to excel on standardized tests are only weakly related to those that determine future CJC, regardless of whether those skills are acquired through good teachers or through other channels. In other words, while high test score effect teachers are largely irrelevant for future CJC, this may be because test scores themselves are largely irrelevant after accounting for confounds.

The observational relationship between behavioral outcomes and future CJC shows a similar pattern. The unconditional relationship suggests large impacts that are attenuated significantly when controlling for family fixed effects among twins. Estimated impacts of high quality teachers as measured by their impacts on behaviors are similar to effects of behaviors themselves within twins. Consistent with previous literature, this suggests that the skills that drive these behaviors are most relevant for CJC. Teachers who successfully develop them consequently have the most meaningful impacts on students.

4.4 Direct effects on long-run outcomes

Table 4 presents variance-covariance estimates for teachers' direct effects on long-run outcomes. As with short-run outcomes, 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 criminal justice interaction: 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 0.028, 0.018, and 0.022 p.p., respectively, or 11.7, 16.3, and 24.4 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. Moreover, teachers' effects on the likelihood of future arrest are positively correlated with their effects on the probability of incarceration.

4.5 Multivariate relationships between teacher effects

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

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

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 5 presents estimates of $\boldsymbol{\beta}$. Consistent with estimates of the correlation structure between short-run outcomes, "horse-racing" the short-run effects changes results little. Test score effects continue to have negligible (although slightly smaller) impacts on future CJC, while behavioral effects have much larger ones. For 12th grade GPA and college attendance, both effects matter independently and enter with meaningful regression coefficients.

Given estimates of the total variation in effects on long-run outcomes from Table 4, it is straightforward to calculate the implied R^2 from these regressions. For criminal arrests, the R^2 is 0.048, while for college attendance it is 0.018. 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.

4.6 Validating effects

4.6.1 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 $\hat{\alpha}_{it}$.[22] 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 3 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.[23] 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 ($\sigma^y = 0.028$—see Table 4) may be biased by $0.028 \cdot 0.00135 = 0.000038$ due to omitted variables. Similar results hold for test scores, behaviors, and academic long-run outcomes (see Appendix Figure A.4 and Appendix 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 Appendix Table A.5 demonstrate that these omitted variables strongly predict test scores, behavioral measures, and study skills.[24] Appendix 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 0.0897 to

[22] 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.

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

[24] Columns 2, 4, and 6 report the regression coefficients underlying Figure 3.

0.108, 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 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).

4.6.2 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 $Var(\eta_j) = 0$. 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.[25]

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 $sg(i, t)$.

[25] 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.

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.[26]

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.[27] 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 6 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 0.998 (0.0142). 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.[28]

The appendix contains several variations on Table 6 that probe the robustness of these results. Appendix 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. Appendix Table A.9 shows the sensitivity of our estimates of λ for our primary long-run outcome, any criminal arrests, with increasingly fine-grained sets of fixed effects, and demonstrates that the instrument is uncorrelated with predicted outcomes based on parental education and twice-lagged test scores.[29] Appendix

[26] 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.

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

[28] 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.

[29] 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. Appendix Table A.9 shows that this holds for CJC.

Table A.10 reports similar results of the same exercise for an important long-run academic outcome, college attendance plans.

4.6.3 Specification robustness

To explore how sensitive our results are to the choice of controls, we estimate a large number of specifications varying them (811 different options) and 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 third-degree polynomials in math and reading scores at least interacted with grade and 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 Appendix Figure A.5 show that our preferred specification is not an outlier. Moreover, to verify that the pre-testing procedure described in Section 3.1 is not driving any of our results, we also replicate Table 5 without it. Appendix Table A.11 shows estimates change little.

5 The role of schools

Variance estimates reflect both teacher variation within schools and cross-school teacher sorting. Understanding the distribution of teacher quality across and within schools is important for policies that seek to improve it. Schools may also have important independent impacts on students' outcomes (Cullen, Jacob and Levitt, 2006; Billings, Deming and Rockoff, 2013; Beuermann et al., 2018; Abdulkadiroğlu et al., 2020; Jackson et al., 2020; Bruhn, 2020; Billings, Deming and Ross, 2019). If school effects are not captured by our controls, some portion of our estimates may actually reflect the causal effects of schools instead of teachers. This section explores these questions.

5.1 Do teachers sort across schools?

Imagine that there are two schools and four teachers. If the teachers in school A have test score effects of 0.5 and those in school B have effects of 1, the overall variance of teacher effects is 0.083. However, within a given school, there are *no* differences in teacher quality. Understanding the degree to which teacher effects are driven by differences within or between schools is important for both the interpretation of our previous results and for

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

policy decisions. In this simple two school example, improving a student's teacher quality requires a move across schools.

The variance of teacher effects can be decomposed using the law of total variance into the average within-school variance and a component reflecting sorting of teachers across schools:

$$Var(\alpha_j) = \underbrace{E_s[Var(\alpha_j|s)]}_{\text{Avg. within-school variance of teacher effects}} + \underbrace{Var_s\left(E[\mu_j|s]\right)}_{\text{Variation in avg. teacher effects between-schools}} \tag{11}$$

Since teachers often switch schools, here we define $Var(\alpha_j)$ as the year-weighted variance of teacher effects. Expectations in Equation 11 are taken over years, with each teacher assigned to a single school in each year by construction.

Figure 4 presents the the results of this exercise. Sorting across schools explains about 80 percent of the variation in teacher effects on both academic and CJC long-run outcome. However, teachers' short-run quality measures exhibit more within-school variation in teacher effects. For example, 80 percent of the variation in teachers' test score value-added is explained by within-school differences.

Sorting is thus an important feature of the data. Exposing students to teachers with stronger effects on many outcomes, including behaviors, may require exposing them to teachers in different schools entirely. Moreover, retention policies that seek to replace bad teacher may, in effect, amount to replacing large portions of the staff at particular schools. Importantly, teacher sorting does not invalidate our design. It does, however, raise the question of whether omitted school effects may explain some of our results, a question we investigate next.

5.2 Are teacher effects actually school effects?

Since much of the variation in teacher effects occurs between rather than within schools, it is important to examine how accounting for schools explicitly would change our results. We conduct two exercises to do so. First, we estimate the variance-covariance of teacher effects on all outcomes separately for each school using Equation 4 and report the 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.

Table 7 reports the implied regression coefficients summarizing the relationship between short- and long-run effects using this approach. Estimates are very similar to those using both within- and across-school variation in teacher effects in Table 5. 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.

The second exercise estimates the variance-covariance of teacher effects using teachers who switch schools and exploiting 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 test scores in school A also do so in school B, or whether teachers who improve behaviors in school A reduce CJC in school B, etc.

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

$$\left(\frac{J-1}{J}\right)\frac{1}{J}\sum_{j=1}^{J}\binom{S_j}{2}^{-1}\sum_{s=1}^{S_j-1}\sum_{k=s+1}^{S_j}\bar{Y}_{js}\bar{Y}_{jk}-2\cdot\frac{1}{J^2}\cdot\sum_{j=1}^{J-1}\sum_{k>j}^{J}\bar{Y}_j\bar{Y}_k \tag{12}$$

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 12 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.

Appendix Table A.12 reports estimates of infeasible regressions of long-run 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 that of estimates that utilize all variation. Although long-run academic outcomes are slightly more sensitive to focusing on teachers switches, behavioral effects continue to be positively related to college attendance and 12th grade GPA.

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 2.2 p.p. standard deviation relative to 2.8 in estimates leaving out a year rather than a school. 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. As in the primary estimates, however, all short-run effects continue to explain a relatively small share of the variance in long-run effects (<7 percent). Omitted school effects are therefore unlikely to explain our core conclusions.

6 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. Teachers' impacts on particular types of students and in particular schools may 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.

We focus on four heterogeneity factors: white vs. non-white students, boys vs. girls, students eligible for free lunch vs. ineligible, and student with above vs. below median predicted arrest risk. We focus on these four dimensions for two reasons. First, inequality along sex, race, and socioeconomic dimensions has been the focus of numerous studies. Second, there are large differences in the likelihoods of interacting with the criminal justice system along these dimensions.

Appendix Table A.13 presents means of student characteristics and average 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 children: seven p.p. (33 percent) lower likelihood of a criminal arrest and 4.6 p.p. (55 percent) lower likelihood of being incarcerated. Similarly, girls and children from higher socioeconomic backgrounds have lower CJC rates. Interestingly, inequalities in long-run academic outcomes are not always correlated with inequalities in CJC. For example, white and non-white students have similar rates of college attendance (46 and 44 percent).

Importantly, the largest inequality in both CJC and long-run academic outcomes is with respect to our socioeconomic indicator, eligibility for free or reduced-price lunch. Poorer students are 13 p.p. (81 percent) more likely to be arrested and 7.7 p.p. (179 percent) to be incarcerated. Moreover, gaps in academic outcomes are also present. Poorer students are less likely to graduate high school (10 p.p., 11 percent) and attend college (35 p.p., 74 percent). These extreme disparities highlight the importance of poverty in predicting adult interactions with the criminal justice system and disparities in human capital accumulation.

6.1 Demographic heterogeneity

Figure 5 summarizes the heterogeneity in teacher effects for short- and long-run outcomes by plotting the estimated correlation in teacher effects across groups. 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 eligible for free lunch vs. ineligible, 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 5 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 on long-run outcomes (e.g., Gershenson et al., 2018). These teachers may be better at promoting skills unmeasured by short-run outcomes for these students.

Finally, we connect these estimates by calculating the implied effects on long-run outcomes of exposing students to teachers with higher student-specific quality. Since short-run effects are so similar across groups, increasing student-specific quality largely implies exposing students in both group 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 (see Appendix Figure A.6).

Still, Figure 6 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 children eligible for free lunch. 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.

6.2 School-level heterogeneity

Motivated by the results in Section 5, we also explore the possibility of heterogeneity in teacher effects based on where the teacher works. It is possible that teachers adjust the skills they focus on in the classroom in response to the school environment. To explore this question, we adjust the model to allow for heterogeneity in teacher effects by school:

$$Y_{ijt} = \underbrace{\sum_j \left(\alpha_j + H'_{s(i,t)}\beta_j\right) D_{ijt}}_{\text{School specific teacher effects } (\alpha_{js})} + X'_{it}\Gamma + \epsilon_{it} \tag{13}$$

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. We denote the effect of teacher j at school s as α_{js}.

Appendix Table A.14 reports estimates of the implied regression coefficients when allowing for heterogeneity in teacher effects by school. As in Table 7, the results are consistent with those assuming a constant effect. Again, among short-run measures of teacher quality the strongest predictor of future CJC is behaviors. Effects on test scores have a negligible impact.

7 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 (e.g., Biasi, 2021). Ideally, a district would evaluate teachers based on their impacts on students' long-run well being (e.g., college attendance, CJC, earnings). 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 connect our our variance estimates to relevant quantities for these simulations, throughout this section we assume that teacher effects are normally distributed.[31]

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:

$$\text{Index}_j^{\text{long-run}} = \omega \mu_j^C + (1-\omega)\mu_j^A \tag{14}$$

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 Panel (a) of Figure 7 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 6 p.p. for exposed students. Naturally, increasing one outcome requires reducing effects on the other. Where a district would locate on this frontier would depend on their preferences over long-run outcomes.

Since teachers' effects on long-run outcomes are not observed, these estimates represent the upper bound of improvements that any policy that seeks to improve these outcomes 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 7 demonstrates feasible gains from doing so. We construct a weighted index of teacher effects on study skills, behaviors, and test

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

scores:

$$\text{Index}_j^{\text{short-run}} = \omega_1 \mu_j^T + \omega_2 \mu_j^B + (1 - \omega_1 - \omega_2)\mu_j^S \tag{15}$$

We then examine 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 Panel (a) of Figure 7 reports the results of these simulations. 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 feasible 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., ones 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).[32] These results are shown in the solid orange line in Panel (a) of Figure 7. 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 and approach, the cost implies reductions in arrests or improvements in college attendance that are roughly half as large.

[32] See Appendix C for full implementation details.

Panel (b) of Figure 7 shows that the trade-offs between improving college attendance and high school graduation are similar to those documented in Panel (a). Panel (b) also demonstrates that the importance of teacher effects on behaviors is not limited to CJC. Regardless of whether the policymaker cares only about college attendance or high school graduation, she should evaluate teachers using both their impacts on students' test scores and behaviors.

8 Conclusion

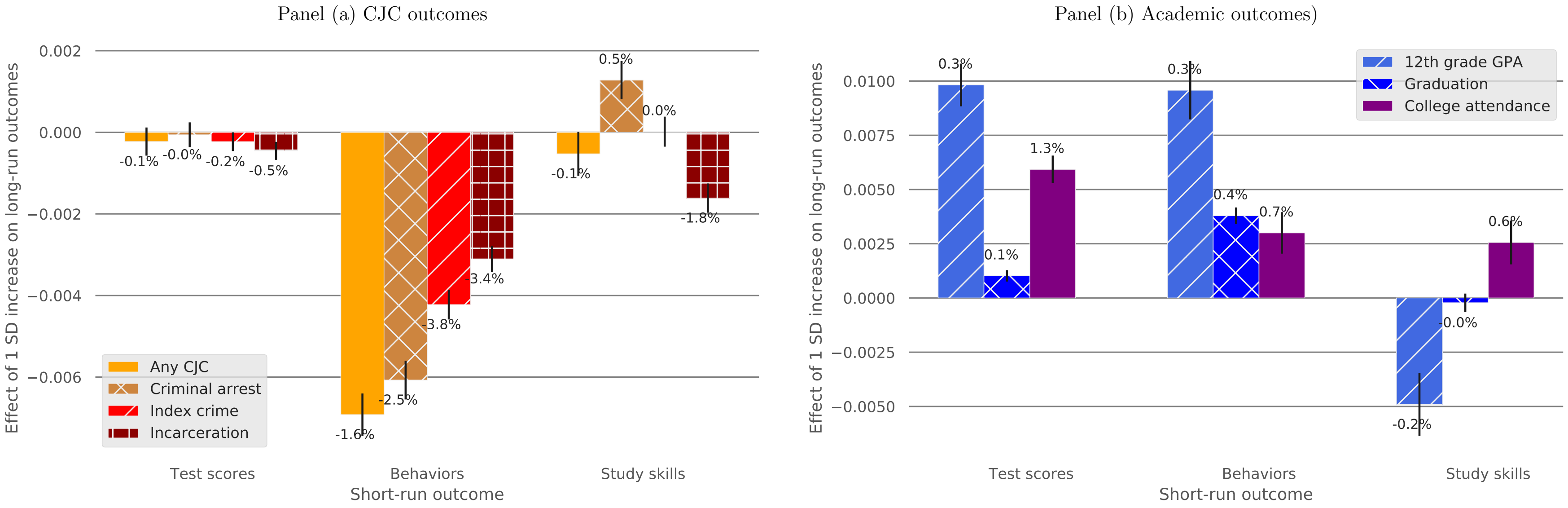
Teachers help students develop a variety of skills necessary to be successful, healthy, and happy adults. The skills needed to excel in one aspect of life, such as the labor market, may differ from those needed in another, such as avoiding entanglement in the criminal justice system. Although prior work demonstrates that teachers who increase students' cognitive skills captured by standardized tests scores increase their propensity to attend college and their 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 plausible 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.

Figures

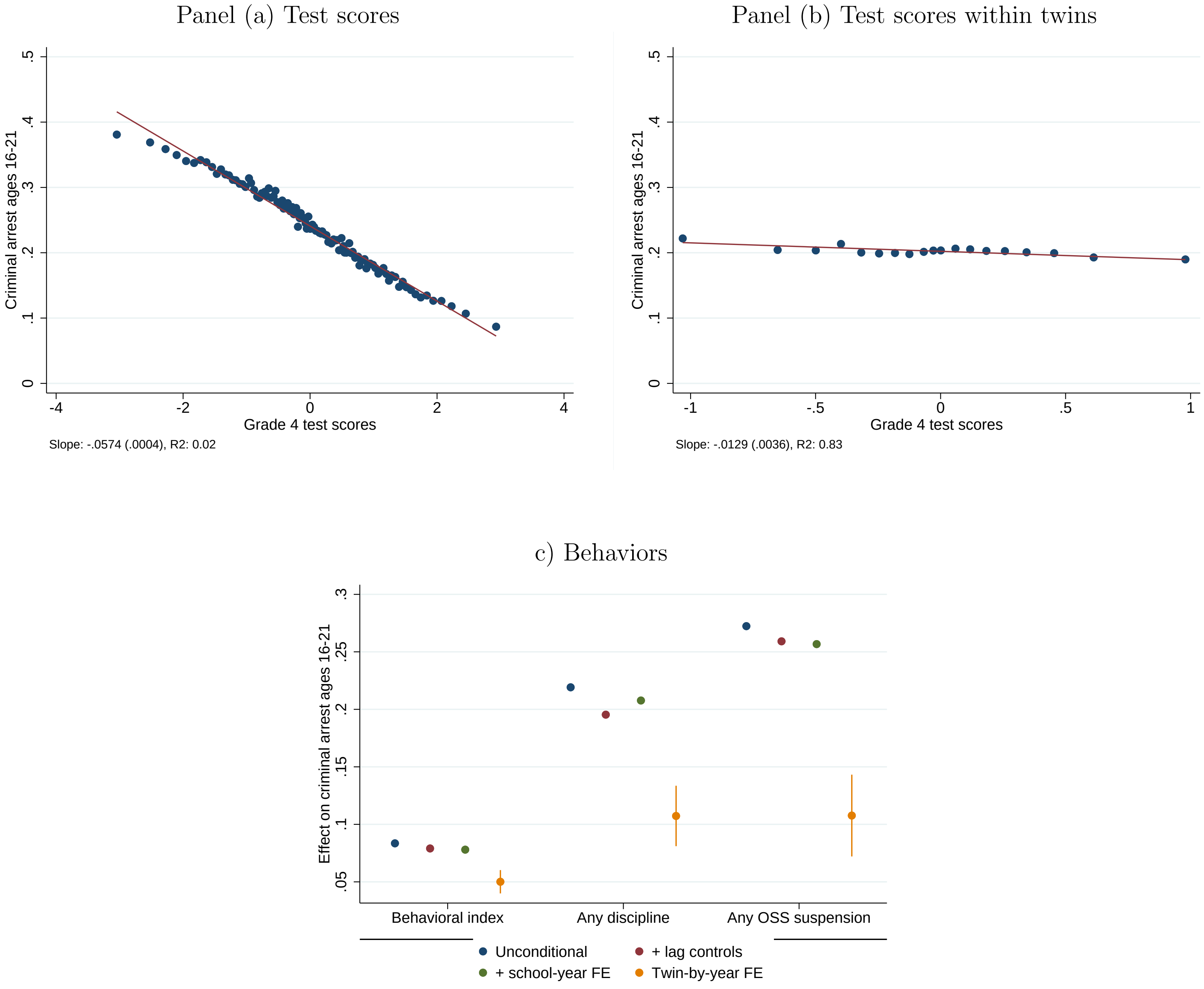
Figure 1: Effects of teacher quality on long-run outcomes


Panel (a) CJC outcomes

Panel (b) Academic 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 calculated using standard errors derived from a weighted bootstrap described in Appendix B. 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: Observational effects of test scores and behaviors on CJC



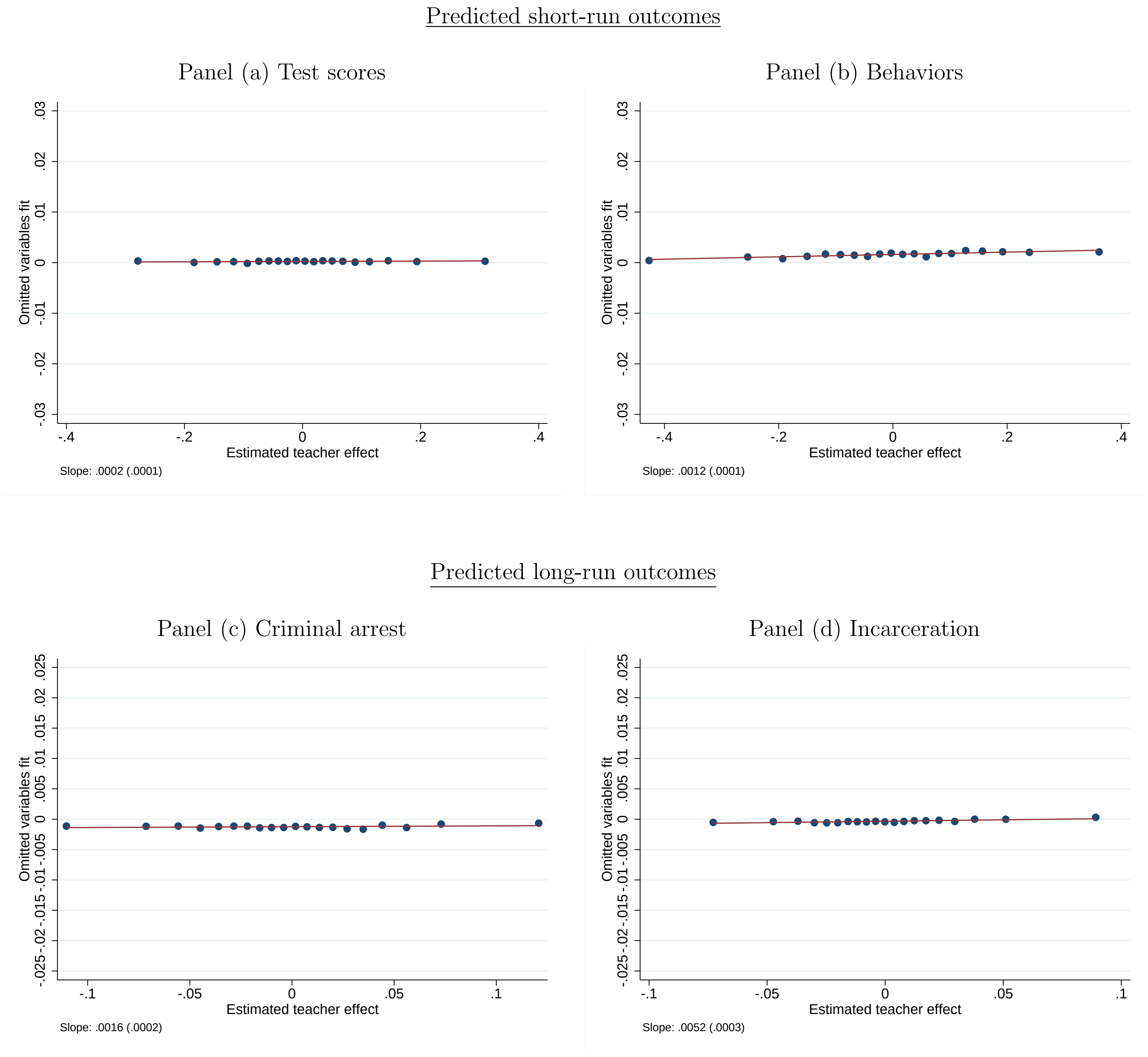
Panel (a) Test scores

Panel (b) Test scores within twins

c) Behaviors

Notes: This figure presents the correlation between grade 4 test scores (Panels a and b) and multiple behavioral measures (Panel c) and an indicator for any criminal arrest between the ages of 16 and 21. Panel a includes all the individuals in our sample and plots the raw correlation without any controls. Panel b includes only twins and controls for twin-by-year fixed effects, so that effects are estimated using within-family variation only. The unconditional slope in the full sample (Panel a) is -0.0574 (0.0004) and when using within-family variation using the twins sample the slope is -.0129 (.0036). Panel c includes the unconditional relationship, effects including lagged outcomes, effects with school-by-year fixed effects, and effects among twins with twin-by-year fixed effects. The estimates include all data where the outcome is observed.

Figure 3: Assessing omitted variable bias in teacher effect estimates



Slope: .0002 (.0001)

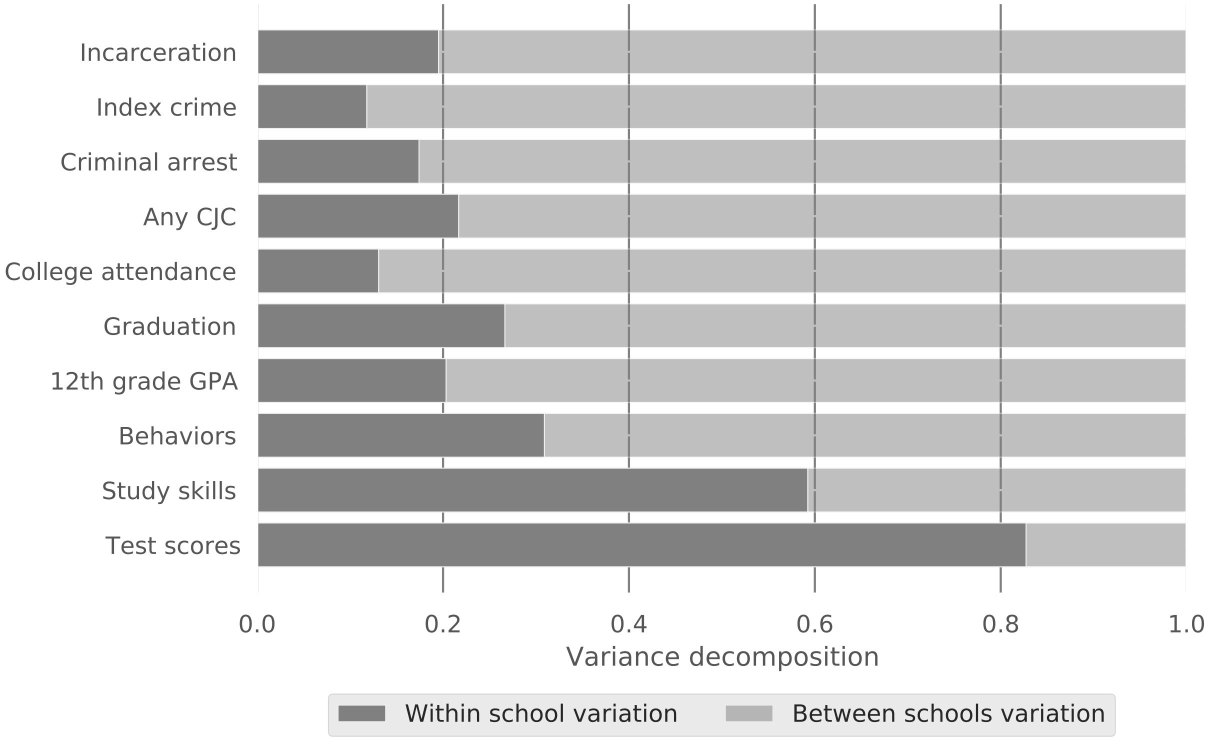
Slope: .0012 (.0001)

Slope: .0016 (.0002)

Slope: .0052 (.0003)

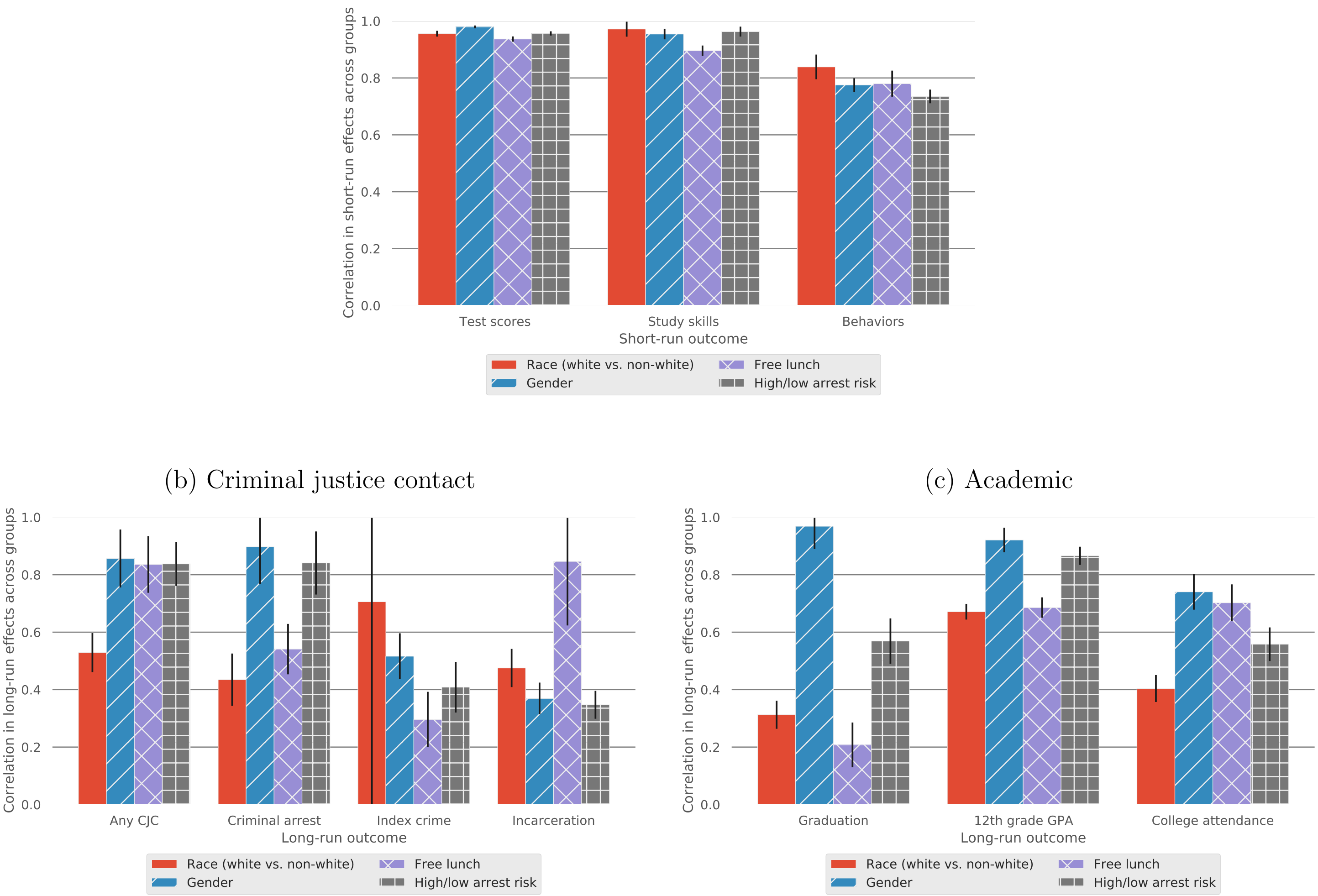
Notes: This figure presents a diagnostic test for whether the estimated teacher effects ($\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$ from Equation 2) are correlated with variables ($W'_{it}\hat{\rho}$ from Equation 9) that predict short- and long-run outcomes but were omitted when estimating 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 4: Between and within-school components of teacher effect variances



Notes: This figure presents a decomposition of the year-weighted variance in teacher effects into a within-school component and a between-school component according to Equation 11. 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: Correlation in teacher effects across groups



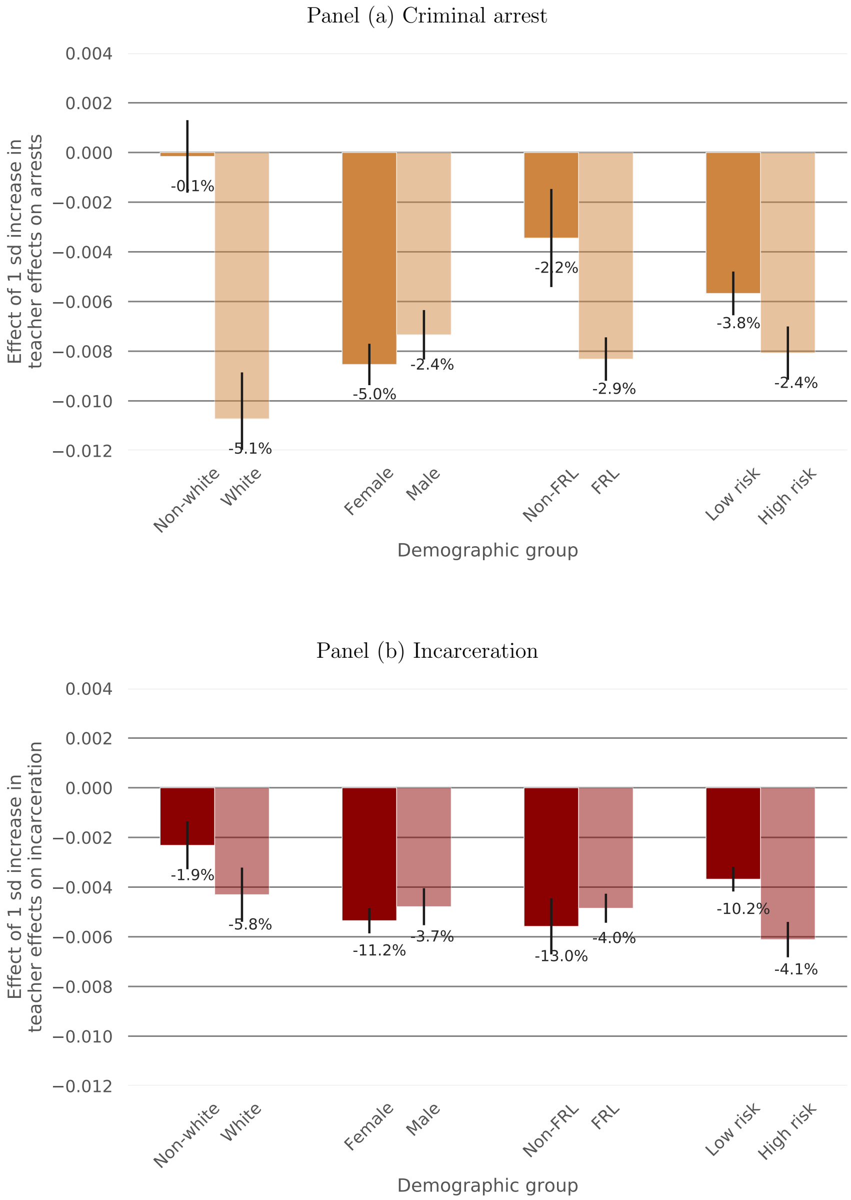
(a) Short-run outcomes

(b) Criminal justice contact

(c) Academic

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 calculated using standard errors derived from a weighted bootstrap described in Appendix B. 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.

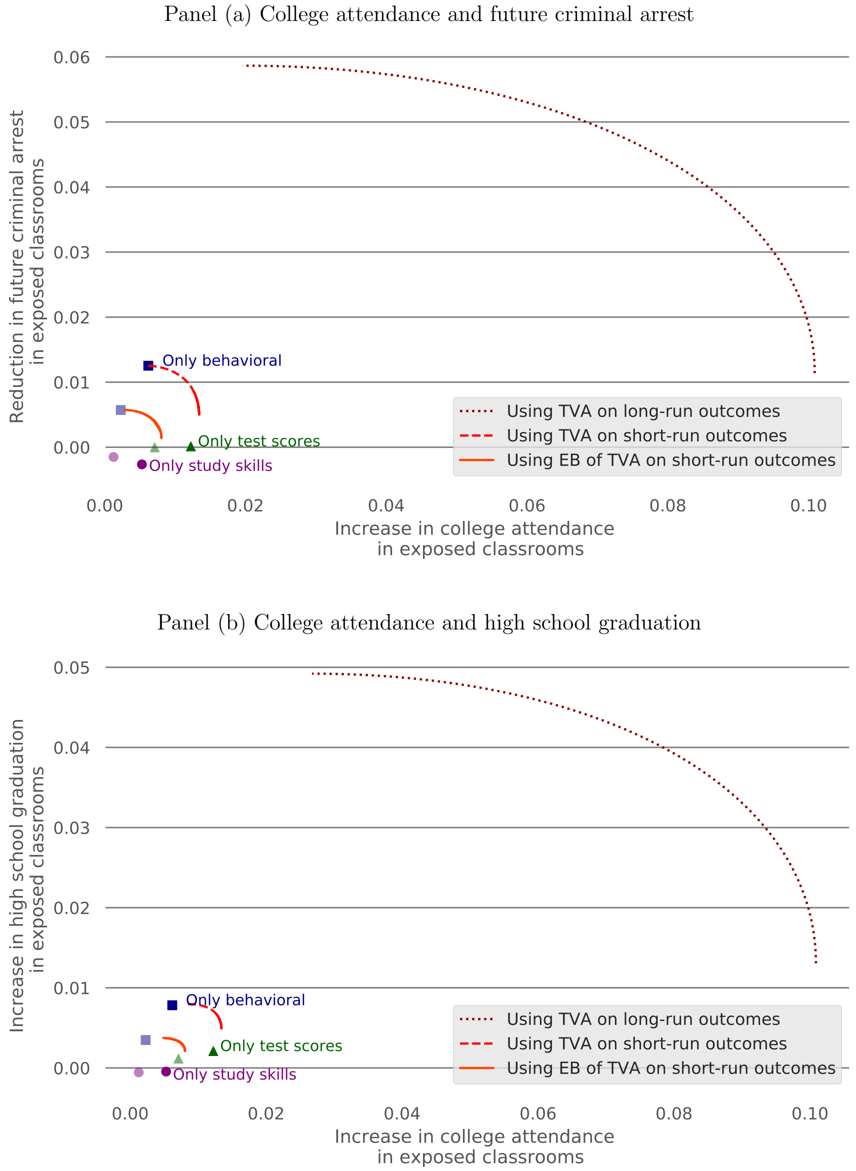
Figure 6: Heterogeneous impacts of exposure to teachers who improve behaviors


Panel (a) Criminal arrest

Panel (b) Incarceration

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 calculated using standard errors derived from a weighted bootstrap described in Appendix B. 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 7: Effects of teacher removal policies on exposed students



Panel (a) College attendance and future criminal arrest

Panel (b) College attendance and high school graduation

Notes: This figure presents simulations of the impacts of replacing the bottom five percent of teachers with an average teacher on college attendance, future criminal arrests, and high school graduation. The rightmost dotted maroon lines in each panel reflect the frontiers achievable if teachers true long-run effects were directly observed. The dashed red line reflects possibilities if teachers true short-run effects on test scores, behaviors, and study skills were observed. The leftmost solid lines shows possibilities using empirical Bayes estimates of teacher effects on short-run outcomes. 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.

Tables

Table 1: Summary statistics

	Full sample		Students w/ CJC	
	Mean	SD	Mean	SD
Demographics				
Male	0.50	0.50	0.64	0.48
Black	0.25	0.43	0.36	0.48
Receives free / subsidized lunch	0.59	0.49	0.74	0.44
Limited English	0.042	0.20	0.035	0.18
Parents have HS education or less	0.40	0.49	0.53	0.50
Parents have some college	0.44	0.50	0.37	0.48
Parents have 4-year degree	0.22	0.41	0.15	0.36
Short-run outcomes				
Standardized reading scores	0.038	0.97	-0.23	0.94
Standardized math scores	0.049	0.96	-0.21	0.92
Days absent	9.03	9.06	10.6	10.8
Any discipline	0.16	0.37	0.29	0.45
Any out-of-school suspension	0.077	0.27	0.19	0.39
Repeat grade	0.0088	0.093	0.015	0.12
Behavioral index	0	1.09	0.43	1.37
Time spent on homework	0.021	0.99	-0.072	1.01
Time spent reading	0.0047	0.99	-0.13	0.96
Time spent watching TV	-0.0065	0.98	0.083	1.02
Study skills index	0	1.09	-0.16	1.09
Long-run outcomes				
12th grade GPA (0-6 scale)	3.12	0.95	2.63	0.87
12th grade class rank	0.49	0.29	0.61	0.26
Graduate high school	0.91	0.28	0.80	0.40
Plans to attend 4-year college	0.45	0.50	0.33	0.47
Any CJC 16-21	0.44	0.50	1	0
Traffic infraction	0.33	0.47	0.63	0.48
Criminal arrest	0.24	0.43	1	0
Index crime arrest	0.11	0.31	0.44	0.50
Criminal conviction	0.10	0.30	0.43	0.49
Incarcerated (jail or prison)	0.090	0.29	0.36	0.48
N student-subject-years	7,422,917		735,270	
N teachers	33,880		23,160	
N students	1,776,759		165,853	
N twin pairs	17,413		3,872	

Notes: This table presents summary statistics for demographic characteristics, short-run outcomes, and long-run outcomes for the analysis sample 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 (62 percent of the analysis sample).

Table 2: Teacher effects on short-run outcomes

	Test scores	Math scores	Reading scores	Study skills	Behaviors
Test scores	0.123 (0.000)	0.912 (0.001)	0.815 (0.002)	0.314 (0.004)	0.067 (0.004)
Math scores		0.136 (0.000)	0.681 (0.003)	0.280 (0.003)	0.046 (0.004)
Reading scores			0.075 (0.000)	0.323 (0.005)	0.101 (0.005)
Study skills				0.190 (0.001)	0.044 (0.004)
Behaviors					0.129 (0.001)

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects on short-run outcomes. Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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 3: Correlation between short- and long-run teacher effects

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	College attendance	Graduation
Test scores	-0.006 (0.005)	-0.002 (0.005)	-0.013 (0.006)	-0.021 (0.005)	0.087 (0.004)	0.121 (0.007)	0.043 (0.006)
Behaviors	-0.193 (0.007)	-0.214 (0.008)	-0.233 (0.010)	-0.143 (0.007)	0.085 (0.006)	0.061 (0.010)	0.159 (0.008)
Study skills	-0.015 (0.008)	0.045 (0.008)	0.001 (0.010)	-0.074 (0.008)	-0.044 (0.007)	0.052 (0.011)	-0.009 (0.009)

Notes: This table presents estimated correlation between teachers short- and long-run effects. Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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 4: Direct effects on long-run outcomes

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	College attendance	Graduation
Any CJC	0.036 (0.000)	0.767 (0.008)	0.537 (0.014)	0.501 (0.010)	-0.034 (0.009)	-0.113 (0.015)	-0.199 (0.013)
Criminal arrest		0.028 (0.000)	0.853 (0.010)	0.608 (0.010)	-0.161 (0.010)	-0.194 (0.017)	-0.340 (0.014)
Index crime			0.018 (0.000)	0.574 (0.012)	-0.148 (0.012)	-0.137 (0.019)	-0.396 (0.015)
Incarceration				0.022 (0.000)	-0.085 (0.009)	-0.176 (0.015)	-0.345 (0.013)
12th grade GPA					0.113 (0.001)	0.423 (0.014)	0.327 (0.012)
College attendance						0.049 (0.001)	0.265 (0.018)
Graduation							0.024 (0.000)

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects on long-run outcomes. Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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 5: 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	0.003 (0.001)	-0.001 (0.001)	-0.000 (0.001)	0.002 (0.001)	0.098 (0.004)	0.008 (0.001)	0.045 (0.003)
Behaviors	-0.058 (0.002)	-0.052 (0.002)	-0.036 (0.002)	-0.026 (0.001)	0.078 (0.006)	0.031 (0.002)	0.023 (0.004)
Study skills	-0.002 (0.001)	0.008 (0.001)	0.001 (0.001)	-0.008 (0.001)	-0.048 (0.004)	-0.004 (0.001)	0.004 (0.003)
$sd(\mu_j^y)$	0.036 (0.000)	0.028 (0.000)	0.018 (0.000)	0.022 (0.000)	0.113 (0.001)	0.024 (0.000)	0.049 (0.001)
R^2	0.037	0.048	0.054	0.026	0.020	0.027	0.018

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. Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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 6: 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
$\hat{\alpha}_{it}$	0.998	1.035	0.920	0.967	1.063	1.068
	(0.0142)	(0.0181)	(0.0842)	(0.155)	(0.0484)	(0.0720)
Observations	7422917	7422917	4058162	4058162	2595879	2595879
R2	0.753	0.751	0.246	0.244	0.180	0.177
Design controls	✓	✓	✓	✓	✓	✓
First stage F	58960	58487	4929	4908	415	413
P-value for $\lambda = 1$	.907	.053	.343	.833	.193	.343

(b) Long-run outcomes

	Criminal arrest		Incarceration		College attendance	
	(1) Schl-grd	(2) Schl	(3) Schl-grd	(4) Schl	(5) Schl-grd	(6) Schl
$\hat{\alpha}_{it}$	1.064	1.327	1.093	1.290	0.976	0.781
	(0.107)	(0.315)	(0.115)	(0.385)	(0.0897)	(0.236)
Observations	3102156	3102156	3102156	3102156	2376836	2376836
R2	0.0924	0.0895	0.0851	0.0825	0.280	0.277
Design controls	✓	✓	✓	✓	✓	✓
First stage F	567	568	180	180	100650	2189
P-value for $\lambda = 1$	.55	.298	.42	.452	.785	.355

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 7: Implied regression of long-run effects on short-run effects using within-school variance-covariance estimates

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	-0.003 (0.002)	-0.007 (0.002)	-0.004 (0.001)	-0.002 (0.001)	0.068 (0.006)	0.007 (0.002)	0.046 (0.004)
Behaviors	-0.048 (0.008)	-0.047 (0.007)	-0.018 (0.005)	-0.029 (0.005)	-0.046 (0.021)	0.002 (0.006)	0.001 (0.015)
Study skills	-0.003 (0.003)	0.007 (0.003)	-0.001 (0.002)	-0.003 (0.002)	0.003 (0.007)	0.002 (0.002)	0.003 (0.005)
$sd(\mu_j^y)$	0.016 (0.000)	0.012 (0.000)	0.006 (0.000)	0.009 (0.000)	0.051 (0.001)	0.012 (0.000)	0.018 (0.001)
R^2	0.038	0.075	0.045	0.043	0.026	0.007	0.086

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. Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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.

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.*

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.

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.

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.

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.

Jacob, Brian A, Lars Lefgren, and David P Sims. 2010. “The persistence of teacher-induced learning.” *Journal of Human resources*, 45(4): 915–943.

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.

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. S. Machin E. Hanushek and L. Woessmann. Amsterdam:Elsevier Science.

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. Elsevier.

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.

Preschool education, educational attainment, and crime prevention: Contributions of cognitive and non-cognitive skills. 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.

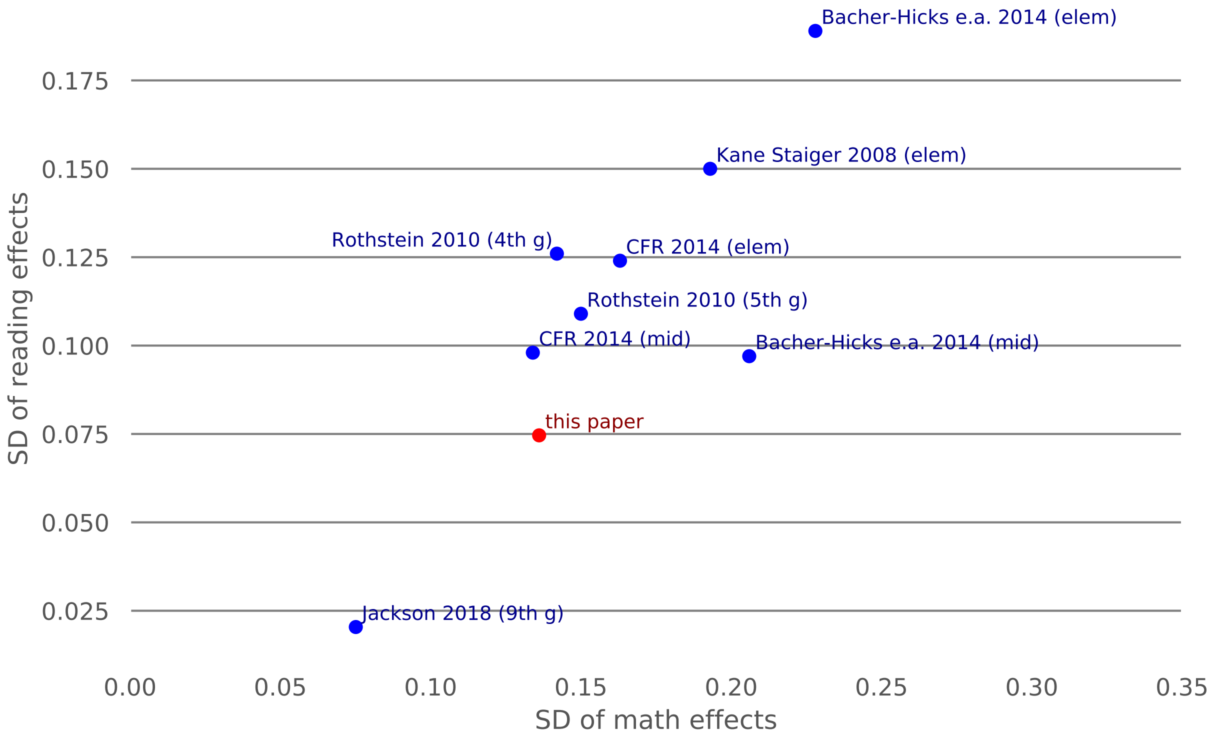
Rubin, Donald B. 1981. "The Bayesian Bootstrap." *The Annals of Statistics*, 130–134.

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.

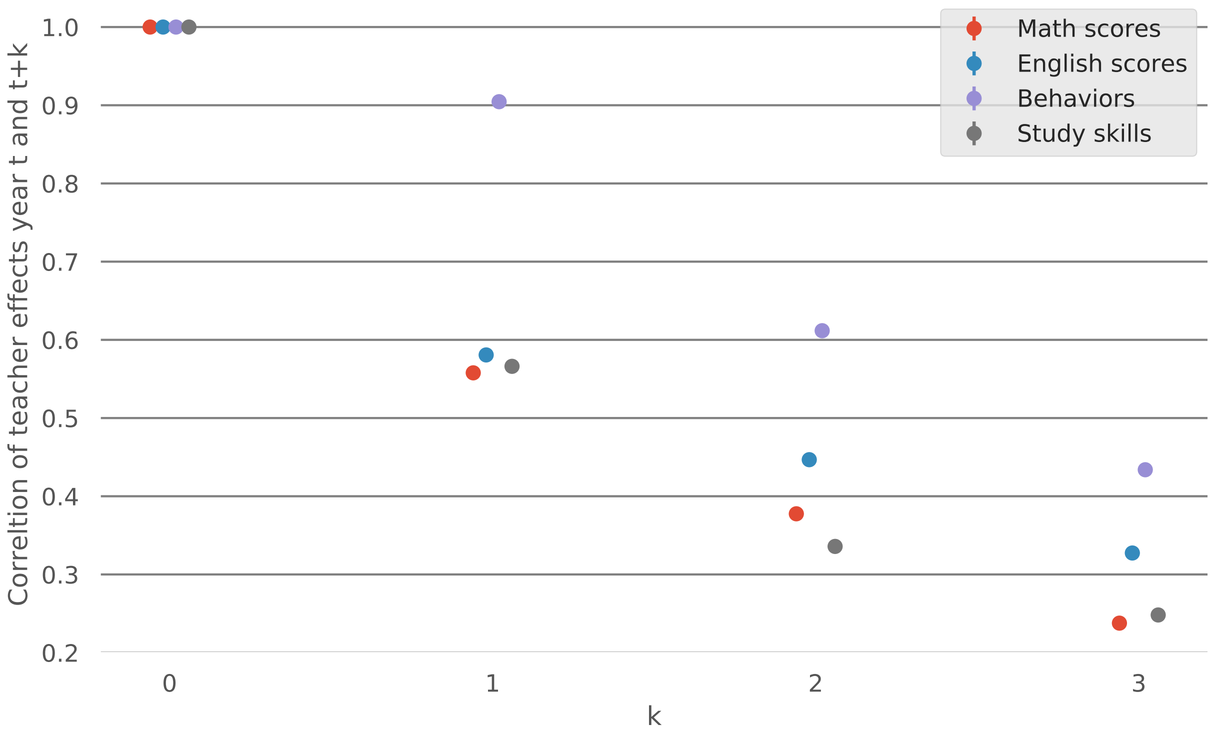
A Additional figures and tables

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



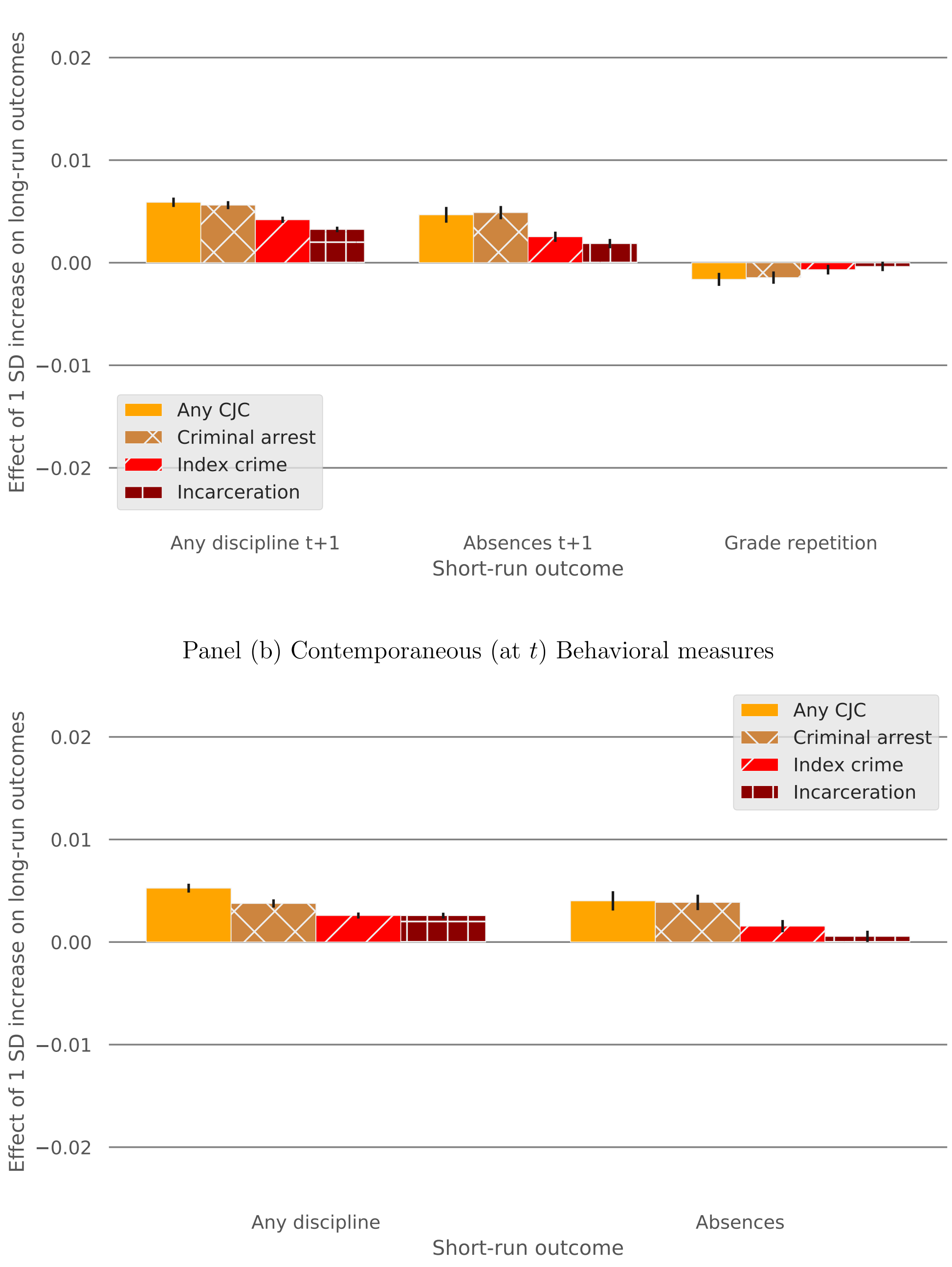
Notes: This figure compares estimated standard deviation of teacher effects on math and reading scores to comparable estimates in the literature. “Mid” indicates estimates for middle school students and “elem” indicates elementary school students. Our estimates straddle those from studies that focus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

Figure A.2: Fade-out of short-run teacher effects



Notes: This table presents estimated correlation in teacher effects on short-run outcomes in year t with effects on outcomes in year $t+1$. The error bars are 95% confidence intervals calculated using standard errors derived from a weighted bootstrap described in Appendix B. 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. 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: Long-run effects of specific behavioral quality dimensions

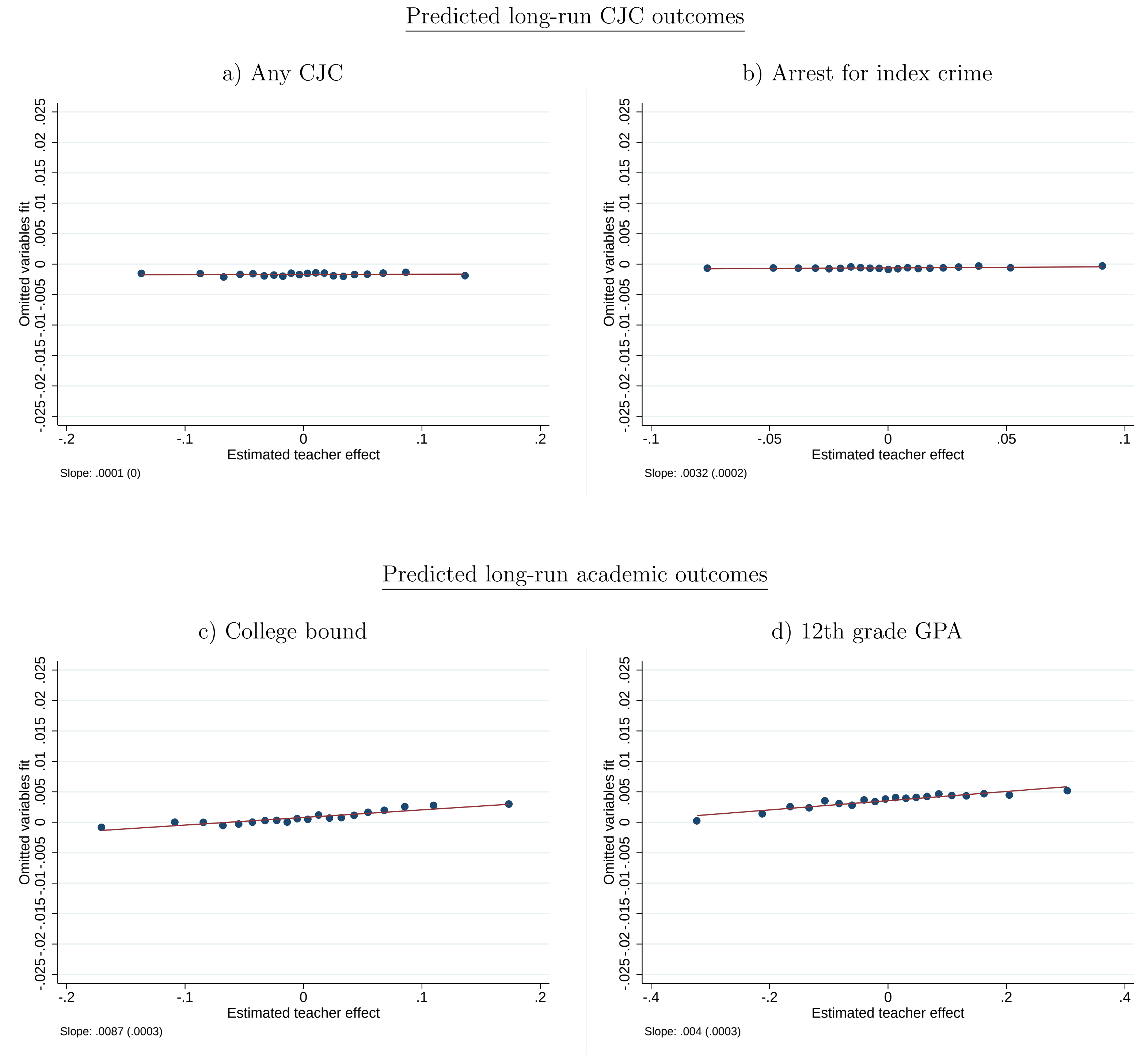


Panel (a) Behavioral measures at $t + 1$

Panel (b) Contemporaneous (at t) 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 calculated using standard errors derived from a weighted bootstrap described in Appendix B. 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.4: Assessing omitted variable bias in additional long-run outcomes


Predicted long-run CJC outcomes

a) Any CJC

b) Arrest for index crime

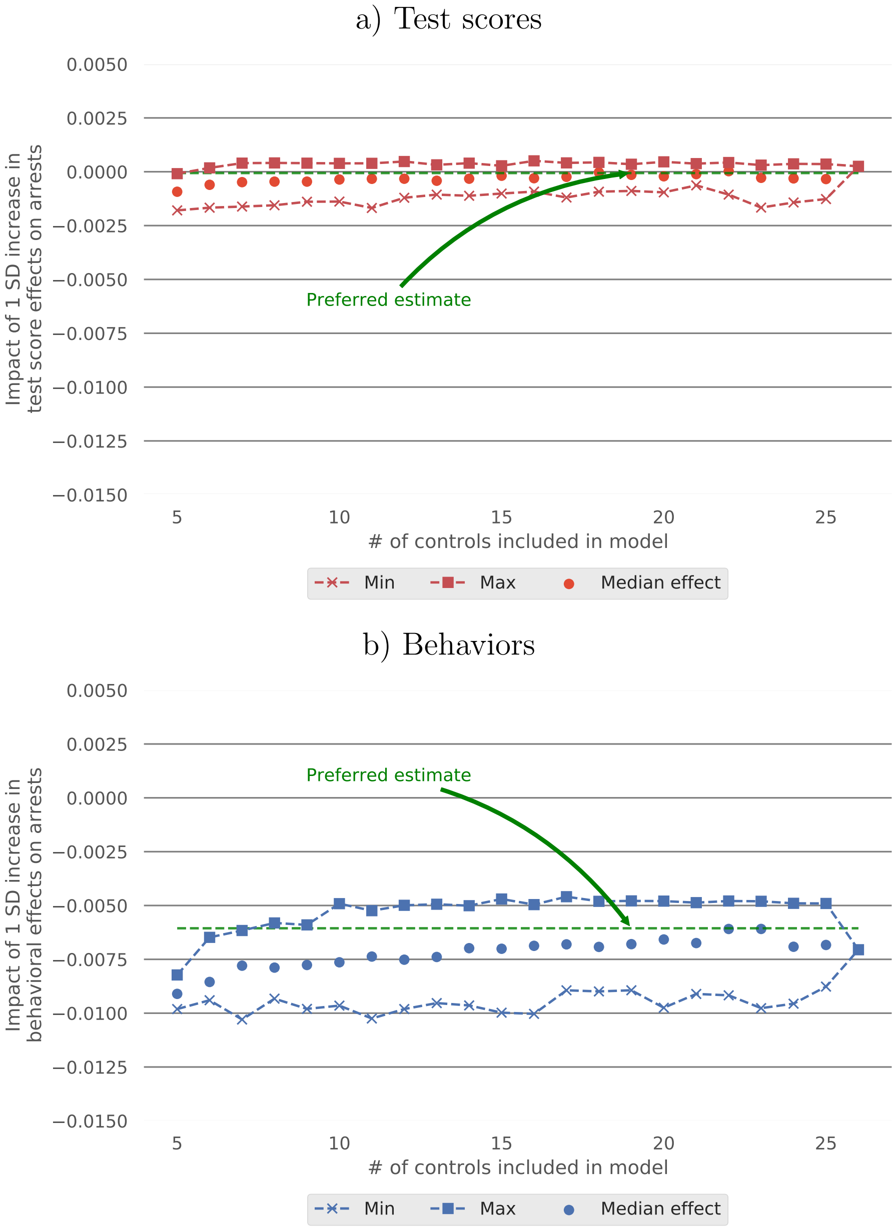
Predicted long-run academic outcomes

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.5: Specification sensitivity of teacher quality impacts on arrests

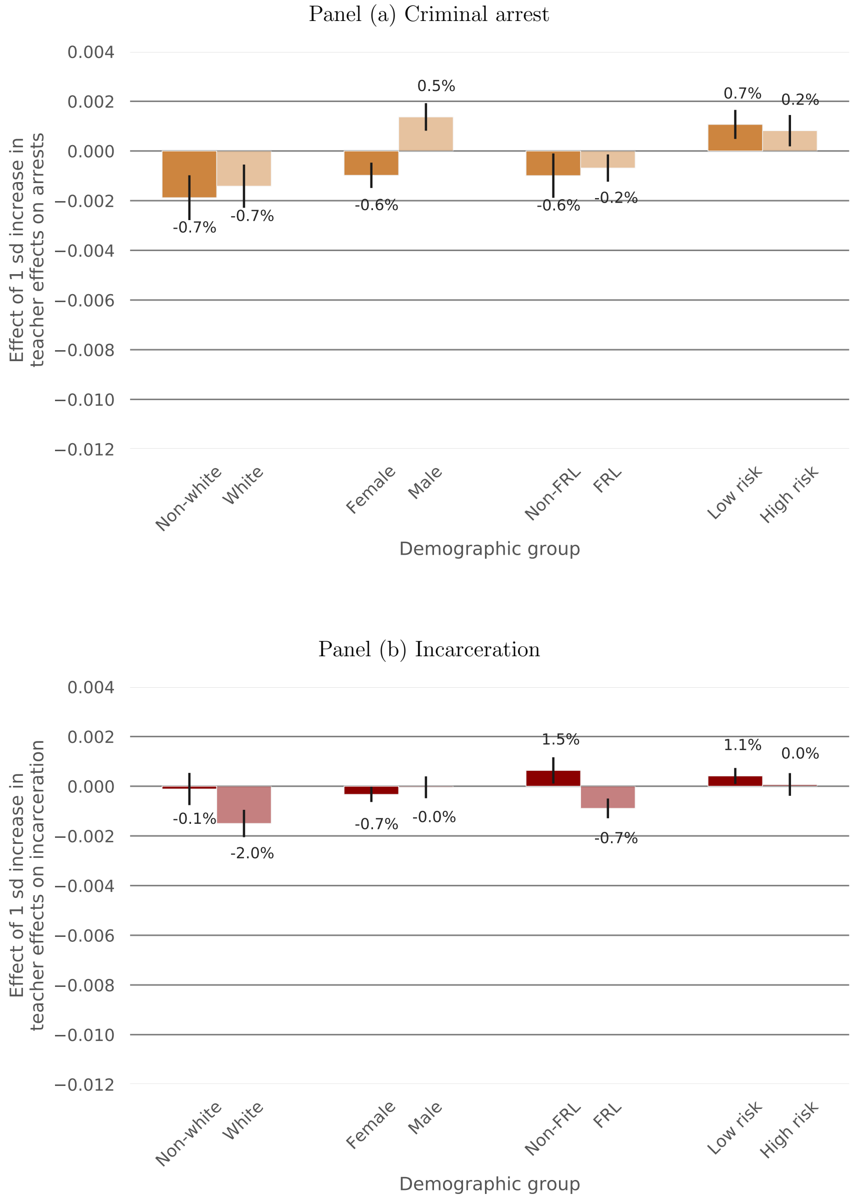


a) Test scores

b) Behaviors

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 at least 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



Panel (a) Criminal arrest

Panel (b) Incarceration

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 calculated using standard errors derived from a weighted bootstrap described in Appendix B. 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 $t+1$	0.060 (0.000)	0.487 (0.006)	0.040 (0.007)	0.076 (0.008)
Any OSS $t+1$		0.026 (0.000)	0.008 (0.011)	0.224 (0.011)
Grade repetition			0.008 (0.000)	-0.017 (0.012)
Days absent $t+1$				0.756 (0.007)

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects on behavioral proxies for non-cognitive skills. Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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: Illustrating of sorting pre-test

a) Unsorted school: χ^2 p-value ≈ 0.92

4th grade teacher	5th grade teacher			**Total**
	D	E	F	
A	11	9	10	30
B	10	8	10	28
C	9	12	10	31
Total	30	29	30	

b) Sorted school: χ^2 p-value ≈ 0

4th grade teacher	5th grade teacher			**Total**
	D	E	F	
A	2	22	1	25
B	1	3	28	32
C	25	3	1	29
Total	28	28	30	

Notes: This table illustrates the pre-test used to identify school-grades where teacher assignments predict teacher assignments in prior grades. For each school-grade, we construct the contingency table of current and lag teacher assignments. We then collect the χ^2 test statistics and p-values for a test of independence of rows and columns. A small p-value is interpreted as evidence that students are tracked to teachers. The main estimates exclude the 20 percent of school grades with p-values below 0.1.

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	-0.00971	-0.00848	-0.00697	-0.00729	0.0801	0.00928	0.0382
	(0.00452)	(0.00380)	(0.00261)	(0.00263)	(0.0101)	(0.00251)	(0.00531)
Design controls	✓	✓	✓	✓	✓	✓	✓
1SD effect	-.0011	-.0009	-.0008	-.0008	.0087	.001	.0042
R2	0.0473	0.0744	0.0592	0.0659	0.532	0.103	0.249
Observations	3102156	3102156	3102156	3102156	2548031	3433225	2376836

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.

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

	CJC outcomes				Academic outcomes		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College bound
Behavioral index VA	-0.0625	-0.0570	-0.0435	-0.0329	0.178	0.0274	0.0415
	(0.00752)	(0.00643)	(0.00444)	(0.00436)	(0.0151)	(0.00407)	(0.00866)
Design controls	✓	✓	✓	✓	✓	✓	✓
1SD effect	-.0049	-.0045	-.0034	-.0026	.014	.0021	.0033
R2	0.0471	0.0741	0.0592	0.0647	0.533	0.102	0.252
Observations	2666011	2666011	2666011	2666011	2133660	2872296	2049307

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.

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		-0.0520	
	(0.00108)		(0.00313)		(0.00290)	
High school only	-0.0340		-0.0726		-0.0220	
	(0.000912)		(0.00220)		(0.00270)	
Some college	0.0172		0.0505		0.0418	
	(0.000948)		(0.00241)		(0.00275)	
BA or more	0.0272		0.0293		0.0613	
	(0.000755)		(0.00229)		(0.00196)	
Lag 2 math	0.115		0.0297		0.00251	
	(0.000473)		(0.00114)		(0.00161)	
Lag 2 reading	0.123		0.0170		0.0562	
	(0.000453)		(0.00109)		(0.00152)	
Teacher effect		0.000339		0.00232		0.00157
		(0.000213)		(0.000506)		(0.000546)
Observations	7406293	7406293	4052845	4052845	2583452	2583452
R2	0.760	0.809	0.251	0.925	0.190	0.796
Original R2	.7495		.2405		.173	
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 even 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 odd 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)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Y	$\hat{Y}$	Y	$\hat{Y}$	Y	$\hat{Y}$	Y	$\hat{Y}$
No high school	0.00349		0.0187		0.0226		0.0283	
	(0.00130)		(0.00109)		(0.000795)		(0.000739)	
High school only	-0.0000594		0.0190		0.0152		0.0181	
	(0.00100)		(0.000846)		(0.000615)		(0.000571)	
Some college	-0.00381		-0.0174		-0.0152		-0.0172	
	(0.00106)		(0.000893)		(0.000649)		(0.000603)	
BA or more	-0.0203		-0.0183		-0.00887		-0.00630	
	(0.000909)		(0.000767)		(0.000557)		(0.000518)	
Lag 2 math	0.0119		0.00316		-0.000218		0.000857	
	(0.000665)		(0.000561)		(0.000408)		(0.000379)	
Lag 2 reading	-0.00913		-0.00780		-0.00528		-0.00509	
	(0.000622)		(0.000525)		(0.000381)		(0.000355)	
Teacher effect		0.000344		0.00135		0.00193		0.00463
		(0.00107)		(0.00108)		(0.00122)		(0.00126)
Observations	3091482	3091482	3091482	3091482	3091482	3091482	3091482	3091482
R2	0.0823	0.795	0.108	0.805	0.0931	0.717	0.102	0.740
Original R2	.0631		.0897		.0739		.0821	
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 even 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 odd 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	-0.0514		-0.0841		-0.0264	
	(0.00239)		(0.000758)		(0.00166)	
High school only	-0.0831		-0.0385		-0.0516	
	(0.00158)		(0.000571)		(0.00109)	
Some college	0.0235		0.0352		0.0270	
	(0.00162)		(0.000594)		(0.00112)	
BA or more	0.183		0.00301		0.120	
	(0.00130)		(0.000479)		(0.000926)	
Lag 2 math	0.103		0.00899		0.0303	
	(0.000967)		(0.000347)		(0.000673)	
Lag 2 reading	0.0449		0.00619		0.0214	
	(0.000918)		(0.000327)		(0.000637)	
Teacher effect		0.00758		0.0213		0.0125
		(0.000857)		(0.00127)		(0.00104)
Observations	2538330	2538330	3420701	3420701	2368241	2368241
R2	0.579	0.853	0.154	0.796	0.305	0.429
Original R2	.5567		.1217		.2743	
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 even 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 odd 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) Schl-grd	(2) Schl	(3) Schl-grd	(4) Schl	(5) Schl-grd	(6) Schl
$\hat{\alpha}_{it}$	0.998	1.033	1.293	1.275	1.138	1.099
	(0.0158)	(0.0201)	(0.284)	(0.481)	(0.100)	(0.112)
Observations	7422917	7422917	4058162	4058162	2595879	2595879
R2	0.722	0.721	0.205	0.203	0.133	0.130
Design controls	✓	✓	✓	✓	✓	✓
School-grade FE	✓	✓	✓	✓	✓	✓
First stage F	32927	32764	957	959	415	414
P-value for $\lambda = 1$	.899	.101	.303	.568	.167	.376

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 CJC

	Outcome: Y			Outcome: $\hat{Y}_{excluded}$		
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\alpha}_{it}$	1.064	1.436	1.409			
	(0.107)	(0.646)	(0.651)			
Z_{it}				0.000949	0.00199	0.00304
				(0.000903)	(0.000916)	(0.000946)
Observations	3102156	3102156	3102156	7422917	7422917	7422917
R2	0.0924	0.0748	0.0731	0.653	0.659	0.669
Design controls	✓	✓	✓	✓	✓	✓
School-grade FE		✓	✓		✓	✓
Dist-grade-year FE			✓			✓
First stage F	567	174	187			
P-value for $\lambda = 1$	.541	.5	.529			

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: Instrumental variables bias tests for teacher-effects on college attendance

	Outcome: Y			Outcome: $\hat{Y}_{excluded}$		
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\alpha}_{it}$	0.970	-0.608	3.946			
	(0.0872)	(4.275)	(21.00)			
Z_{it}				0.00314	0.000721	0.00243
				(0.00125)	(0.00127)	(0.00131)
Observations	2376836	2376817	2376604	7422917	7422917	7422917
R2	0.280	0.173	0.0521	0.608	0.616	0.628
Design controls	✓	✓	✓	✓	✓	✓
School-grade FE		✓	✓		✓	✓
Dist-grade-year FE			✓			✓
First stage F	4767	441	511			
P-value for $\lambda = 1$	.727	.707	.888			

Notes: This table presents instrumental variable tests for bias in estimated teacher effects on plans to attend college in the future. 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.11: Implied regression of long-run effects on short-run effects without pre-testing procedure

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	-0.001 (0.001)	-0.003 (0.001)	-0.003 (0.001)	0.000 (0.001)	0.097 (0.004)	0.010 (0.001)	0.045 (0.003)
Behaviors	-0.059 (0.002)	-0.048 (0.002)	-0.038 (0.001)	-0.023 (0.001)	0.124 (0.006)	0.039 (0.002)	0.029 (0.004)
Study skills	-0.004 (0.001)	0.007 (0.001)	0.002 (0.001)	-0.009 (0.001)	-0.014 (0.004)	0.001 (0.001)	0.009 (0.003)
$sd(\mu_j^y)$	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 (0.001)
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. The tables is analogous to Table 5 but without conducting the pre-testing procedure that identifies school-grades where students' lagged teacher assignments predict current teacher assignments. Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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.12: 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	0.003 (0.002)	-0.020 (0.001)	-0.004 (0.001)	0.005 (0.001)	-0.089 (0.006)	-0.010 (0.002)	-0.002 (0.004)
Behaviors	-0.057 (0.004)	-0.029 (0.004)	-0.025 (0.003)	-0.021 (0.003)	0.181 (0.016)	-0.009 (0.004)	0.104 (0.010)
Study skills	-0.001 (0.002)	0.025 (0.001)	-0.001 (0.001)	-0.006 (0.001)	0.092 (0.005)	0.023 (0.001)	0.026 (0.003)
$sd(\mu_j^y)$	0.026 (0.000)	0.022 (0.000)	0.008 (0.001)	0.016 (0.000)	0.073 (0.001)	0.016 (0.000)	0.032 (0.001)
R^2	0.025	0.040	0.066	0.011	0.061	0.048	0.068

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 5. Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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.13: Summary statistics of four different sub-groups comparisons across race, sex, socioeconomic status, and predicted risk of arrest

	Full sample	Race		Sex		Free lunch		Arrest risk	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		White	Non-White	Boys	Girls	Eligible	Not eligible	High	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	0.24	0.25	0.36	0.085	0.40	0.10
Receives free / subsidized lunch	0.59	0.43	0.81	0.58	0.59	1	0	0.78	0.39
Limited English	0.042	0.0036	0.094	0.043	0.041	0.067	0.0073	0.016	0.069
Parents have HS education or less	0.40	0.33	0.50	0.40	0.41	0.56	0.19	0.57	0.24
Parents have some college	0.44	0.52	0.33	0.44	0.44	0.26	0.68	0.30	0.59
Parents have 4-year degree	0.22	0.28	0.13	0.22	0.21	0.062	0.43	0.092	0.34
Short-run outcomes									
Standardized reading scores	0.038	0.28	-0.29	-0.027	0.10	-0.27	0.47	-0.34	0.41
Standardized math scores	0.049	0.28	-0.26	0.051	0.046	-0.26	0.49	-0.31	0.41
Days absent	9.03	9.49	8.42	9.19	8.87	9.99	7.47	10.5	7.71
Any discipline	0.16	0.12	0.22	0.22	0.10	0.21	0.076	0.27	0.072
Any out-of-school suspension	0.077	0.043	0.12	0.11	0.044	0.11	0.025	0.14	0.025
Repeat grade	0.0088	0.0062	0.012	0.011	0.0063	0.013	0.0028	0.015	0.0027
Behavioral index	-1.7e-09	-0.063	0.085	0.13	-0.13	0.19	-0.32	0.34	-0.25
Time spent on homework	0.021	0.086	-0.071	-0.013	0.055	-0.082	0.16	-0.11	0.13
Time spent reading	0.0047	0.045	-0.057	-0.13	0.14	-0.052	0.083	-0.16	0.14
Time spent watching TV	-0.0065	-0.15	0.20	0.063	-0.076	0.14	-0.18	0.15	-0.20
Study skills index	7.2e-11	0.11	-0.17	-0.13	0.13	-0.15	0.20	-0.22	0.29
Long-run outcomes									
12th grade GPA (0-6 scale)	3.12	3.33	2.79	2.95	3.27	2.78	3.53	2.59	3.52
12th grade class rank	0.49	0.44	0.55	0.54	0.44	0.57	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	0.45	0.46	0.44	0.41	0.50	0.34	0.59	0.31	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	0.38	0.29
Criminal arrest	0.24	0.21	0.28	0.30	0.17	0.29	0.16	0.33	0.15
Index crime arrest	0.11	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	0.055	0.13	0.052	0.16	0.043
Incarcerated	0.090	0.074	0.12	0.13	0.048	0.12	0.043	0.15	0.036
N student-subject-years	7,422,917	4,247,740	3,175,161	3,717,275	3,705,641	4,337,242	3,026,396	3,711,459	3,711,458
N teachers	33,880	33,553	33,838	33,877	33,880	33,861	33,549	33,866	33,802
N students	1,776,759	965,547	811,205	893,250	883,508	996,015	738,348	1,065,174	976,201
N twin pairs	17,413	9,682	8,973	11,099	11,408	11,446	7,465	12,414	10,432

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), eligible for free (or price reduced) lunch and non-eligible (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.14: Implied regression of long-run effects on short-run effects using within-school variance-covariance estimates

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	-0.000 (0.002)	-0.008 (0.002)	-0.008 (0.001)	-0.005 (0.001)	0.095 (0.005)	0.010 (0.002)	0.059 (0.004)
Behaviors	-0.084 (0.015)	-0.055 (0.014)	-0.026 (0.011)	-0.028 (0.010)	-0.101 (0.042)	0.003 (0.011)	-0.040 (0.027)
Study skills	-0.006 (0.004)	0.007 (0.003)	0.002 (0.002)	-0.006 (0.002)	-0.017 (0.009)	-0.009 (0.003)	-0.010 (0.007)
$sd(\mu_j^y)$	0.009 (0.000)	0.006 (0.000)	0.004 (0.000)	0.007 (0.000)	0.042 (0.001)	0.012 (0.000)	0.008 (0.001)
R^2	0.317	0.395	0.223	0.095	0.083	0.015	0.871

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 13 that allows teacher effects to vary by school (i.e., α_{js}). Standard errors in parentheses are derived from a weighted bootstrap described in Appendix B. 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 Weighted bootstrap procedure for inference on variance-covariance estimates of teacher effects

To conduct inference on the variance and covariance estimates using Equations 4 and Equation 7, we use the weighted bootstrap procedure proposed by Rubin (1981). This procedure is especially well suited to for our setting. Specifically, according to the model in Equation 2, the only stochastic component is u_{it}. Moreover, the distribution of teachers across schools and students across teachers and schools are both fixed. The objective of our inference procedure is to estimates the variability in our estimates due to sampling error in u_{it} while holding fixed the other components of the model (e.g., the distribution of teachers across schools). The weighted bootstrap of Rubin (1981) uses all the data in each bootstrap iteration but places different weights on different parts to simulate sampling variability.

In each bootstrap sample b:

1. Randomly simulate a vector of weights of length N (the number of teacher-year observations) from an exponential distribution with a rate parameter of one, $\boldsymbol{V}^b \sim exp(1)$. We then normalize the weights $\boldsymbol{V}^b$ to sum to one within each teacher and denote the normalized weights by $\boldsymbol{\omega}^b$.

2. Next we calculate a weighted version of the estimator in Equation 4 (or Equation 7):

$$\widehat{Var}(\alpha_j)^b = \left(\frac{J-1}{J}\right)\frac{1}{J}\sum_{j=1}^{J}\frac{1}{\sum_{t=1}^{T_j-1}\sum_{k=t+1}^{T_j}\omega_{jt}\omega_{jk}}\sum_{t=1}^{T_j-1}\sum_{k=t+1}^{T_j}\omega_{jt}\omega_{jk}\bar{Y}_{jt}\bar{Y}_{jk} - 2\cdot\frac{1}{J^2}\cdot\sum_{j=1}^{J-1}\sum_{k>j}^{J}\left(\sum_{t=1}^{T_j}\omega_{jt}\bar{Y}_{jt}\right)\left(\sum_{t=1}^{T_k}\omega_{kt}\bar{Y}_{kt}\right) \tag{B.1}$$

   We then repeat steps 1 and 2 for B times and our bootstrap estimate of the SE is the standard deviation across bootstrap iterations:

$$\widehat{SE} = SD(\widehat{Var}(\alpha_j)^1, \widehat{Var}(\alpha_j)^2, \ldots, \widehat{Var}(\alpha_j)^B) \tag{B.2}$$

Bootstrapped standard errors for other parameters, such as implied regression coefficients, are estimated analogously.

C Policy simulation details

This appendix includes technical details for the implementation of the simulations discussed in Section 7. 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 7, 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.

C.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 14 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 $(\mu^A, \omega\mu^C + (1-\omega)\mu^A)$.

C.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 15 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 $(\mu^A, \omega_1\mu^T + \omega_2\mu^B + (1-\omega_1-\omega_2)\mu^S)$.

C.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 15.

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:

$$\begin{aligned} \text{Index}_j^{\text{EB short-run}} &= E[\underbrace{\omega_1 \mu^T + \omega_2 \mu^B + (1 - \omega_1 - \omega_2)\mu^S}_{=\text{Index}_j^{\text{Ideal short-run}}} | \hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S] && \text{(C.1)} \\ &= E[\text{Index}_j^{\text{Ideal short-run}}] + \boldsymbol{b}_I' \left[(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S) - E[(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)] \right] && \text{(C.2)} \end{aligned}$$

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., $\Sigma^{-1}_{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}} \Sigma_{\hat{\boldsymbol{\mu}}\text{Index}_j^{\text{EB short-run}}}$ 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 $(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)$:

$$\begin{aligned} \hat{Y}^{\text{EB short-run}} &= E[Y | \hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S] && \text{(C.3)} \\ &= E[Y] + \boldsymbol{b}_Y' \left[(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S) - E[(\hat{\mu}^T, \hat{\mu}^B, \hat{\mu}^S)] \right] && \text{(C.4)} \end{aligned}$$

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:

$$\begin{aligned}
&\underbrace{E[\hat{Y}^{\text{EB short-run}}]}_{=E[Y]} - E[\hat{Y}^{\text{EB short-run}}|\text{Index}_j^{\text{EB short-run}} < q_{0.05}^{\text{EB of short-run}}] \qquad \text{(C.5)}\\
&= \frac{Cov(\hat{Y}^{\text{EB short-run}}, \text{Index}_j^{\text{EB short-run}})}{Var\left(\text{Index}_j^{\text{EB short-run}}\right)} E[\text{Index}_j^{\text{EB short-run}}|\text{Index}_j^{\text{EB short-run}} < q_{0.05}^{\text{EB of short-run}}]\\
&= \frac{Cov(\hat{Y}^{\text{EB short-run}}, \text{Index}_j^{\text{EB short-run}})}{Var\left(\text{Index}_j^{\text{EB short-run}}\right)} \sigma_{\text{Index}_j^{\text{EB short-run}}} \frac{\phi\left(\frac{q_{0.05}^{\text{EB of short-run}} - E[\text{Index}_j^{\text{EB short-run}}]}{\sigma_{\text{Index}_j^{\text{EB short-run}}}}\right)}{\Phi\left(\frac{q_{0.05}^{\text{EB of short-run}} - E[\text{Index}_j^{\text{EB short-run}}]}{\sigma_{\text{Index}_j^{\text{EB short-run}}}}\right)}
\end{aligned}$$

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.