The Effects of Teacher Quality on Adult Criminal Justice Contact

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

September 2025

Abstract

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

*Rose: Associate Professor, University of Chicago and NBER; ekrose@uchicago.edu. Schellenberg: Economist at Amazon Web Services; jschellenberg@econ.berkeley.edu. Shem-Tov: Assistant Professor, University of California, Los Angeles and NBER; shemtov@econ.ucla.edu. We are extremely grateful to Christopher Walters, David Card, Jesse Rothstein, and Patrick Kline, who provided invaluable support and advice. We thank the North Carolina Education Research Data Center for assistance in the construction of our data set and to Justin McCrary and the Spencer Foundation, who helped fund this project. Thanks also go to Natalie Bau, Gordon Dahl, Zarek Brot-Goldberg, Amanda Dahlstrand, 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, Nestor Le Nestour, Elena Pastorino, Avner Shlain, and Isaac Sorkin for their helpful comments and suggestions. A previous version of this paper was entitled "The Effects of Teacher Quality on Criminal Behavior." Software to implement the methods in this paper is available at https://pypi.org/project/ustat-var/.

The social costs of crime and any resulting arrest and incarceration are large. Cohen and Piquero (2009), for example, estimate the value of averting a high-risk youth from future crime at $3.8 to $5.3 million. One way to reduce these costs is through the education system. Little is known, however, about how teachers affect crime, despite their central role in providing education. Even less is known about which skills good teachers foster to reduce their students' future crime, and how those skills differ from those most important for other outcomes (Chetty et al., 2014b; Jackson, 2018; Jackson et al., 2020; Petek and Pope, 2021; Gilraine and Pope, 2021). Since different students may benefit from focusing on different skills and approaches, an important related question is whether some teachers are better at reaching particular types of students (Dee, 2005; Gershenson et al., 2022).

Studying these questions requires estimating and relating teachers' impacts on multiple outcomes and types of students, which can be difficult with noisy estimates of each teacher's effects. The standard approach is to estimate teachers' value-added on each outcome or student type separately while adjusting for sampling error using Empirical Bayes (EB) techniques.[1] Recent work has highlighted potential violations of the parametric assumptions embedded in standard EB approaches (Gilraine et al., 2021; Chen, 2023). Even if the model is correctly specified, however, the distribution of EB estimates does not consistently estimate the distribution of teacher effects. The variance of shrunken value-added estimates can approach zero, for example, even if teachers' true impacts vary widely, and covariances across groups or outcomes can differ in sign and magnitude from the covariances of true effects.

To help overcome these issues, this paper develops new, EB-free approaches for studying multi-dimensional teacher effects. We then apply these tools to nearly two decades of linked administrative schooling and criminal justice records. We estimate the variance of elementary and middle school teachers' effects on their students' future crime as measured by rates of criminal justice contact (CJC) including arrest, conviction, and incarceration. To study the drivers of these effects, we estimate their covariance with teachers' effects on standardized test scores, behavioral proxies for non-cognitive skills, and study skills. Doing so allows us to ask whether teachers who boost test scores, for example, also decrease their students' future crime. After establishing a

[1]Studies using this approach include Jackson (2018) and Petek and Pope (2021) for impacts on test scores and behaviors, and Bates et al. (2022) and Biasi et al. (2021) for impacts on disadvantaged and non-disadvantaged students, among others.

set of baseline, homogeneous effects, we explore the importance of accounting for heterogeneity across student characteristics.

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

Our empirical strategy uses non-parametric estimators of the variance-covariance structure of teacher effects, building on the literature on variance component estimation (e.g., Krueger and Summers, 1988; Aaronson et al., 2007; Kline et al., 2020). The approach separates the problem of estimating the joint *distribution* of teacher effects—the goal of our analysis—from obtaining good estimates of a set of *individual* teacher's effects—the goal of EB and other forecasting techniques. Many parameters sought in the value-added literature, such as the implied impacts of a one standard deviation increase in teacher quality, can be easily recovered using our approach. As is common practice, the analysis relies on a rich set of characteristics available as controls to justify a conditional independence assumption, which is supported by a battery of validation tests.

We find large teacher effects on both short and long-run outcomes. The estimated standard deviation of teacher effects on future arrests, for example, is 2.7 percentage points (p.p.), or 11.3% of the sample mean, and is 2.1 p.p. (23.6%) for incarceration. The pattern of effects across crime types and severity suggest that reductions stem from decreases in criminal behavior as opposed to better evasion of police. We find similarly large effects on long-run academic outcomes studied in previous analyses, such as high school graduation and students' plans to attend college. Estimated impacts on short-run outcomes are similar to recent estimates in other geographies (Kane and Staiger, 2008; Chetty et al., 2014a).

Relating effects on short- and long-run outcomes reveals that teachers who boost test scores do not meaningfully decrease students' future CJC. Shifting a student to

a teacher with a 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 p.p.; we cannot reject zero effect at conventional significance levels. Teachers who boost study skills have similarly limited effects on CJC. High test score effect teachers do, however, improve students' long-run academic outcomes, with estimated impacts on college attendance similar to those in Chetty et al. (2014b).

By contrast, teachers' impacts on behavioral outcomes are closely connected to their impacts on CJC. Assignment to a teacher with a standard deviation more beneficial impact on a summary index of discipline, attendance, and grade repetition decreases the likelihood of future CJC by 2-4%, depending on the outcome. If teacher effects on short-run behaviors are evidence of influence over non-cognitive and socio-emotional skills and traits such as conscientiousness, perseverance, and sociability (Lleras, 2008; Bertrand and Pan, 2013), our results support a growing body of research suggesting that the accumulation of "soft skills" may lie at the heart of the return to education for crime (Heckman and Rubinstein, 2001; Heckman et al., 2006; Reynolds et al., 2010; Heckman and Kautz, 2012; Heckman et al., 2013; Jackson et al., 2020).[2]

While the estimates rely on non-experimental teacher assignments, multiple tests demonstrate that all effects are measured with limited bias. The estimates are insensitive to the inclusion of covariates unlikely to be used by administrators assigning teachers to students, including parental education, twin indicators, and twice-lagged test scores, all of which strongly predict outcomes conditional on the baseline controls. Using teacher switches across schools and school-grades to instrument for teacher assignments, we cannot reject that estimates are unbiased. Nevertheless, we also show that our estimates provide a lower bound on the variance of causal effects and that covariances in causal effects are still identified under certain forms of bias, including if estimates of test score effects are unbiased but direct effects on CJC are not.

Teachers have substantive impacts on future CJC for many types of students, including groups defined by sex, race, socio-economic status, and predicted CJC risk using all covariates. But effects are not perfectly correlated across student types. The correlation of a teacher's effects on their white and non-white students' criminal

[2] A large literature documents the importance of non-cognitive and socio-emotional skills for long-run outcomes, including Heckman and Rubinstein (2001), Cunha and Heckman (2008), Cunha et al. (2010), Lindqvist and Vestman (2011), Deming (2017), and Gray-Lobe et al. (2023).

arrests is roughly 0.5, for example, indicating important heterogeneity in teachers' impacts. Effects on short-run outcomes, on the other hand, show tight correlation across groups. The correlation between a teacher's test score effects for boys and girls is 0.96, for example. The impact of assignment to a teacher good at boosting test scores or improving behaviors for observably similar students is thus similar to the average effect, since heterogeneous effects on short-run outcomes provide a poor proxy for heterogeneous effects on long-run outcomes.

To conclude, we simulate the impacts of replacing the bottom 5% 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 6 p.p. reductions in criminal arrests for exposed students. Policies that target teachers using their impacts on short-run measures, however, achieve a fraction of these gains, underscoring the scope of teacher impacts not captured by these measures. Putting emphasis on different short-run outcomes trades off effects on long-run academic and CJC outcomes, with policies that emphasize test score impacts most strongly affecting academic outcomes and policies that emphasize behaviors primarily influencing criminal justice outcomes.

This paper contributes to a broad literature on the importance of teachers. Related work has shown that teacher quality—as measured by their influence on students' test scores—varies widely (Rivkin et al., 2005; Kane and Staiger, 2008; Chetty et al., 2014a; Bacher-Hicks et al., 2014; Bau and Das, 2020) and has important consequences (Chetty et al., 2014b). Teachers, however, impact a broad set of skills beyond those measured by standardized tests (Jackson, 2018; Petek and Pope, 2021; Mulhern and Opper, 2022). The skills rewarded in one domain, such as the labor market, may differ from those rewarded in another, implying that what makes a teacher "good" depends on the outcome considered. Our results provide the first evidence of teachers' effects on a life-altering and socially costly long-run outcome and show that the set of short-run impacts that predict these effects is strikingly different than for teacher effects on adult earnings, complementing related work on the impacts of overall school quality (Jackson et al., 2020; Bacher-Hicks et al., 2019; Jordan et al., 2024).

We also make progress methodologically by providing new tools for studying the impacts of teachers. While recent work has suggested a variety of alternatives to traditional EB estimators (e.g., Kwon, 2023; Chen, 2023; Giacomini et al., 2023), the

results of any forecasting technique will depend simultaneously on how much teachers impact outcomes and the researcher's ability to accurately estimate those impacts. In standard approaches, for example, the impacts of a one standard deviation increase in teacher quality would change if estimated in the same data but dropping half of each teacher's observations, making it difficult to reach firm conclusions about the relative importance of teachers' impacts across outcome dimensions and contexts. By directly estimating the parameters of interest under weak assumptions, our approach sidesteps these issues. Building on recent work by Kline et al. (2020), we also derive analytic standard errors that avoid the complications of inference on generated regressors in traditional two-step procedures. A Python implementation of our method is available at https://pypi.org/project/ustat-var/.

Our work also connects to the literature on how student skills respond to investments (Cunha and Heckman, 2008; Cunha et al., 2010, 2021). That literature typically exploits repeated, noisy measurements of latent skills (e.g., Agostinelli and Wiswall (2025) use sub-tests of the Peabody Individual Achievement Test) and independence restrictions to achieve identification (Schennach, 2016; Cunha et al., 2021). Our focus is similar in that we also seek to estimate distributions of latent parameters and use long-run outcomes to anchor effects on outcomes without a natural scale such as test scores. However, we focus on teachers' direct effects on short- and long-run outcomes, rather than estimating dynamic human capital production functions. As a result, we do not explicitly model achievement and disciplinary outcomes as proxies for latent underlying skill stocks. Our identification arguments and estimation approach do not rely on repeated, contemporaneous measurements, knowledge of investments, or latent factor structures, and require generally weaker restrictions.

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

1 Data and setting

1.1 Education records

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

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' test proctor to link students to their teachers.[3]

1.1.1 The teacher assignment process

Typically, central administrators (e.g., the principal) assign students to teachers seeking to balance class size, student ability, and student needs (e.g., special education status). Our empirical strategy controls for the student characteristics used as potential inputs in this process, such as the student's academically gifted status and behavioral or educational special needs, their lagged academic achievement, as well as lagged school discipline and absences. A key concern is whether assignment also depends on student characteristics we do not observe, leading to potentially biased estimates of teachers' causal effects. We discuss tests of this issue using student characteristics likely not used in the assignment process below.

Because students are not randomly assigned to schools or cohorts, and teacher quality varies across both schools and time, our models will also control for year-grade-subject fixed effects to proxy for the implicit "strata" at which administrators assign teachers to students, and means of student characteristics at the school level to control for between-school sorting. Finally, we include classroom means of student characteristics to account for the possibility that some students are "tracked" into different sets of

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

courses and teachers (e.g., the honors track) (Jackson, 2014).

1.1.2 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 students. Test scores are normalized within each year and grade to have a mean of zero and a standard deviation of one in the full student population. For homeroom teachers, we use the first principal component of math and reading scores as the relevant outcome. Math and reading 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 et al., 2006; Lleras, 2008; Bertrand and Pan, 2013; Petek and Pope, 2021). As in Jackson (2018), we take the first principal component of standardized indicators for school discipline (primarily in- and out-of-school suspensions), days absent, and grade repetition in each year. Prior research documents that these behaviors are strongly associated with important non-cognitive skills and traits (e.g., Duckworth et al., 2007). 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) correspond to more positive values of the index.

The third and final measure uses data on students' time spent on homework, reading, and watching television, which we interpret as proxies for students' study skills and effort. These variables are reported categorically with discretization that changes year-to-year. We convert values to hours using the mid-point of each category and normalize within grade and year. As with behaviors, we then combine the three measures into a single summary index using the first principal component. Although test scores are available over the full panel, behaviors and study skill measures are not. Absences are available for all students beginning in 2004, and disciplinary records begin in 2001 for a subset of the schools and for all schools beginning in 2006. When

[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.

estimating teacher effects on each outcome, we use all data available.

1.2 Criminal justice records

We use administrative information on arrests, charges, and sentencing from the NC 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 NC 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% 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 NC 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]

1.3 Data linking

Education records were linked to criminal justice data on the basis of name and date of birth.[7] Since not all students are arrested as young adults, we do not expect 100% 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 et al. (2019), for example, estimate that 19% of Charlotte-Mecklenburg students are arrested between the ages of 16 and 21, a figure close to our rate of criminal arrest

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

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

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

reported below.[8]

1.4 Sample construction

Following Chetty et al. (2014a), 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% of records) or grades (3.7%) in the same year, since their students are likely only partially exposed to their potential effects, alternative and special education schools (0.1%), and students with an invalid contemporary or lag math and reading score, which serve as crucial controls (8.4%). 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% of observations).

1.5 Summary statistics

Table 1 presents summary statistics for the final analysis sample. The sample includes 9,779,708 student-subject-year observations for 1,953,547 students with 39,707 different teachers. Roughly 25% of the sample are black—close to the state average, 58% are economically disadvantaged, 22% have a parent with a four-year college degree, while 40% 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 (Columns 1 and 2) the average math and reading test scores are slightly higher (0.061 and 0.046), primarily due to the exclusion of students without a lag score. 17% of the students have some disciplinary infraction in an average year and 8% have an out-of-school suspension.

Contact with the justice system is common. A quarter of the students have a criminal arrest between ages 16 to 21. A substantial share of the children have a serious incident between ages 16 to 21: 10% are arrested for an index crime, 10% are convicted of a crime, and 9% are incarcerated (including both jail and prison). Columns 3 and 4

[8]Other benchmarks include Cook and Kang (2016), who estimate that 6% of the 1987-89 NC 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% 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.

report statistics for the sub-sample of the students in cohorts for whom we observe CJC outcomes. This sample is remarkably similar to our full analysis sample.

Table 1 also reports summary statistics for children who have a criminal arrest between ages 16 and 21 (Columns 5 and 6). These students are more likely to be economically disadvantaged, their parents have less 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.

2 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 et al., 2014a; 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.

2.1 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.[9] Short-run outcomes, such as test scores, vary across t within student. Long-run outcomes do not, but we preserve the t subscript because Y_{ijt} reflect impacts of potential teacher assignment at time t conditional on the student's history as of time t. Let X_{it} denote the student's exogenous observable characteristics. If all potential outcomes were observed, they could be decomposed into the causal effects of teachers and student observed and unobserved heterogeneity using regression:

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

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

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. Since $E[\mu_j] = 0$, μ_j captures teacher j's effects relative to the average teacher. Since Equation 1 is a simple linear projection of potential outcomes onto observables, as written it imposes no additional structure on the nature of treatment effects. In the first part of our analysis, however, we follow the prior literature and assume teacher effects are homogeneous across students, allowing us to write $\epsilon_{ijt} = \epsilon_{it}$ (Chetty et al., 2014a; Angrist et al., 2017).

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

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

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

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

Conditional independence does *not* necessarily require random assignment of students to teachers. Instead, teacher assignments must be uncorrelated with unobserved factors that influence outcomes. This assumption rules out, for example, some teachers being systematically assigned students who are more likely to do well on standardized tests than observationally similar peers regardless of their teacher. But it allows teachers to be assigned students with different observed or unobserved characteristics so long as their influence on outcomes is accounted for by the controls.[10]

[10]This distinction is important in light of Rothstein (2010)'s influential finding that teacher assignments are correlated with observables such as twice-lagged test scores in a subset of the same NC data we use. Rothstein's result is not necessarily inconsistent with Assumption 1.

The consensus in the education literature is that Assumption 1 is a plausible restriction due to the richness of available controls (Bacher-Hicks and Koedel, 2023). As noted in Section 1.1.1, these controls both proxy for the student characteristics administrators use when making teacher assignments and account for student sorting across schools and cohorts. In our case, the controls include year-grade-subject fixed effects; third-order polynomials in lagged math and reading test scores interacted with grade and subject; indicators for the student's academically gifted status, behavioral or educational special needs, economic disadvantage indicators, and English proficiency status; race and gender; lagged school discipline and grade repetition indicators and lagged days absent; and school and classroom means of lagged math and reading test scores and student characteristics. If a variable is missing for a particular student, it is replaced with zero and a dummy variable for missing is included.

We discuss several tests of Assumption 1 in what follows, all of which support its validity. Nevertheless, we also consider a weaker assumption known as "forecast unbiasedness" that allows for a restricted form of bias in teacher effects:

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.[11] Observational effects are therefore unbiased linear predictors of causal effects. Assumption 2 requires that the causal effects of teachers who appear high quality given the students they are actually assigned cannot be systematically over- or under-estimated. This restricted form of bias is plausible if high quality teachers are sometimes assigned students likely to excel no matter what, but also sometimes assigned students who face more challenges.

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

[11] 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 (i.e., projecting μ_j on $\hat{\alpha}_j$ will not yield a coefficient of one). We detail how we overcome this issue when testing this assumption below.

Likewise, Assumptions 1 and 2 can be invoked for the causal and observational effects of teachers on each outcome separately.

2.2 Parameters of interest and estimation

We focus on estimating the variance-covariance matrix of observational teacher 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' impacts relate across outcomes. Because $\boldsymbol{\alpha}_j$ are defined as population projection coefficients, they are easy to consistently estimate using OLS. But these estimates will also be 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 latent effect variances and covariances (Kline et al., 2020). Our approach allows us to non-parametrically characterize the distribution of teacher effects without the use of an intermediate shrinkage step or specifying a complete statistical model.

To begin, define the variance of teacher effects (for a single generic outcome) as:

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

where the second expression follows from expanding the squares and some careful algebraic rearrangement. The key challenge in estimating $Var(\alpha_j)$ is that the plug-in estimator $Var(\hat{\alpha}_j)$ will be biased because $E[\hat{\alpha}_j^2] > E[\hat{\alpha}_j]^2$ by Jensen's inequality. To overcome this issue, we construct unbiased estimates of α_j^2 for each teacher motivated by a simple restriction on teacher-year mean population residuals from Equation 2. Specifically, we assume that for $\bar{u}_{jt} = \frac{1}{n_{jt}}\sum_{i|j(i,t)=j} u_{it}$, where n_{jt} is the number of students assigned to teacher j at time t, the following holds:

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

If Assumption 1 holds, then Assumption 3 can be understood as requiring that any sorting on unobservables is uncorrelated across t for each teacher. This restriction is not imposed by Assumption 1 alone, which simply requires that unobservable sorting be mean zero for each teacher when averaging across all t. Assumption 3 rules out

in addition any "runs" across years in unobserved student quality within teacher. It is straightforward to relax Assumption 3, however, by assuming that it holds across t separated by at least m years: $E[\bar{u}_{jt}\bar{u}_{jt'}] = 0 \; \forall j, |t - t'| \geq m$. This allows for runs that die out after at least m years, a possibility we explore further below.

The key ingredient in our estimator of $Var(\alpha_j)$ is teacher-year mean residuals from OLS estimates of Equation 2:

$$\bar{Y}_{jt} = \frac{1}{n_{jt}} \sum_{i|j(i,t)=j} Y_{it} - X'_{it}\hat{\Gamma} = \alpha_j + \hat{\bar{u}}_{jt}$$

where $\hat{\bar{u}}_{jt} = \bar{u}_{jt} + \bar{X}'_{jt}(\Gamma - \hat{\Gamma})$ for $\bar{X}_{jt} = \frac{1}{n_{jt}} \sum_{i|j(i,t)=j} X_{it}$.[12] Our estimator uses these mean residuals as follows:

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

where T_j is the number of years observed for teacher j and $\bar{Y}_j = \frac{1}{T_j} \sum_{t=1}^{T_j} \bar{Y}_{jt}$. This estimator can be viewed as the second expression in Equation 3 with $\bar{Y}_j$ substituted for α_j and $\binom{T_j}{2}^{-1} \sum_{t=1}^{T_j-1} \sum_{k=t+1}^{T_j} \bar{Y}_{jt}\bar{Y}_{jk}$ substituted for α_j^2. The latter term is the average product of mean residuals across all pairs of years for a given teacher.

Under Assumption 3, $E_{t \neq t'}\left[\bar{Y}_{jt}\bar{Y}_{jt'}\right] = \alpha_j^2$ because each residual consists of α_j plus uncorrelated noise.[13] Related 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., Kane and Staiger, 2008; Chetty et al., 2014a; 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] Population residuals are *estimated* because Γ is not known. As a result, $\bar{X}'_{jt}(\Gamma - \hat{\Gamma})$ appears in $\hat{\bar{u}}_{jt}$. The presence of this term may generate bias if $\bar{X}_{jt}$ is correlated across years within teacher and $\Gamma \neq \hat{\Gamma}$ even under Assumption 3. Our very large sample makes estimation error in Γ small, mitigating this concern. In fact, bias can be estimated because under Assumption 3 for any j and $t \neq t'$, $E[\bar{Y}_{jt}\bar{Y}_{jt'}] - \alpha_j^2 = E[\bar{X}'_{jt}(\Gamma - \hat{\Gamma})(\Gamma - \hat{\Gamma})'\bar{X}_{jt'}] = \bar{X}'_{jt}\Sigma_\Gamma \bar{X}_{jt'}$. For test scores, this evaluates to 10^{-5} when averaged overall all j and t, implying a $< 0.1\%$ upwards bias in our estimated variance. The same residualization approach is standard in related methods. When estimating two-way firm and worker fixed effect models for log wages, for example, covariates such as year dummies are commonly partialled out before estimating the variance of firm effects (Kline et al., 2020).

[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 the expectation using the median value in 200 Monte Carlo simulations. Our estimator uses all possible pairs of years within a teacher instead of simulations.

$\widehat{Var}(\alpha_j)$ 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 et al., 2020):

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

where $\bar{Y} = \frac{1}{J}\sum_{j=1}^{J}\bar{Y}_j$, 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 et al., 2007; Kline et al., 2021).

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

$$\begin{aligned} Cov(\alpha_j^A, \alpha_j^C) &= \frac{1}{J}\sum_{j=1}^{J}\alpha_j^A\alpha_j^C - \left(\frac{1}{J}\sum_{j=1}^{J}\alpha_j^A\right)\left(\frac{1}{J}\sum_{j=1}^{J}\alpha_j^C\right) \\ &= \left(\frac{J-1}{J}\right)\frac{1}{J}\sum_{j=1}^{J}\alpha_j^A\alpha_j^C - 2\frac{1}{J^2}\cdot\sum_{j=1}^{J-1}\sum_{k>j}^{J}\alpha_j^A\alpha_j^C \end{aligned} \tag{6}$$

where the second line again follows from expanding the product of sums and re-arranging. 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 a version of Assumption 3 holds across outcomes. Specifically, we assume that for generic outcomes A and C:

Assumption 4 *Uncorrelated cross-outcome-year residuals:* $E[\bar{u}_{jt}^A\bar{u}_{jt'}^C] = 0 \;\forall j, t \neq t'$

where $\bar{u}_{jt}^k$ are the teacher-year mean residuals from Equation 2 with outcome Y_{it}^k on the left-hand side. Much like $\widehat{Var}(\alpha_j)$, our covariance estimator exploits this restriction by excluding products of mean residuals from the same year:

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

where $\bar{Y}_{jt}^k$ and $\bar{Y}_j^k$ are defined as above but for outcome Y_{it}^k. As above, this estimator can be viewed as Equation 6 with unbiased estimators of $\alpha_j^A\alpha_j^C$ formed from average

cross-year products for each j plugged in. It is also numerically equivalent to taking the covariance of estimated effects across outcomes and subtracting a correction for within-teacher correlated measurement error:

$$\widehat{Cov}(\alpha_j^A, \alpha_j^C) = \frac{1}{J}\sum_{j=1}^{J} \underbrace{\left(\bar{Y}_j^A - \bar{Y}^A\right)\left(\bar{Y}_j^C - \bar{Y}^C\right)}_{\substack{\text{Covariance of} \\ \text{observed } \hat{\alpha}_j^A \text{ and } \hat{\alpha}_j^C}} - \underbrace{\left(1 - \frac{1}{J}\right)\frac{\sigma_j^{AC}}{T_j}}_{\substack{\text{Correction for correlated} \\ \text{sampling variation}}} \tag{8}$$

where $\sigma_j^{AC} = \frac{1}{T_j - 1}\sum_{t=1}^{T_j}\left(Y_{jt}^A - \bar{Y}_j^A\right)\left(Y_{jt}^C - \bar{Y}_j^C\right)$.[15]

Comparison to alternative approaches. The prior literature often uses the covariance of EB posteriors to study how teachers' impacts across multiple outcomes or groups are related. Standard linear EB posteriors typically shrink observational estimates towards the overall mean with weights related to each estimate's sampling variance. The variance of EB posteriors is generally smaller than the variance of latent effects as a result of this shrinkage step, however. This can make it difficult to compare effects across groups or outcomes, since both the variance of latent effects and degree of shrinkage may be changing. In Appendix B, we show that the covariance of EB posteriors across outcomes or student groups also does not identify the covariance in latent true teacher effects. A simple simulation calibrated to the variance-covariance of teacher effects and sampling error in our data illustrates the potential biases—the covariances of EB posteriors across outcomes can be significantly biased and wrongly signed, even with over 1,000 student observations per teacher.

Inference. An additional benefit of our approach is that it is possible to construct analytic expressions for the sampling variances of the estimators in Equations 4 and 7, as well as their sampling co-variances. Appendix C explains how. We use this result to conduct inference instead of relying on bootstrap routines that may be misleading in high-dimensional models (El Karoui and Purdom, 2018). Doing so also allows us to avoid the inferential complications that arise when using EB posteriors as explanatory variables in second-step regressions.

2.2.1 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

[15] So far, we assumed that all outcomes are observed in all the years. In Appendix C, we present a generalization of our estimator that relaxes this requirement.

observational (α_j) and causal (μ_j) teacher effects coincide.

2.2.2 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., 2021). 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.

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 bias in teacher effects on future CJC is orthogonal to teachers' causal effects on test scores. This expression also makes the direction of any potential bias clear. It seems plausible, for example, that the teachers with the most positive causal effects on test scores are assigned the students least likely to be arrested. This pattern would make estimated observational correlations more negative than causal correlations. As we show below, however, we estimate a near zero correlation between teachers' test score and CJC effects, leaving little scope for large negative correlations in causal effects.

3 The causal effects of teachers

This section begins by estimating teacher effects on students' future CJC and other long-run outcomes, as well as on short-run outcomes. We then examine the correlation structure of effects on short- and long-run outcomes. Finally, we conduct multiple tests of Assumptions 1 and 2 to support the causal interpretation of our estimates and demonstrate the robustness of our results to alternative specifications and assumptions.

3.1 Effects on future CJC

Table 2 presents the estimated variance-covariance of structure of teachers' direct effects on long-run outcomes. The diagonal entries are standard deviations and the off-diagonals are correlations. The first four columns show effects on four measures

of CJC: any interaction (including traffic tickets and other non-criminal violations), criminal arrests, arrests for index crimes, and incarceration. The final three columns show effects on 12th grade GPA (measured on a six point scale), graduation, and plans for four-year college attendance.

Teacher effects on future CJC are large. A one standard deviation increase in effects would increase the likelihood of future criminal arrest, arrests for index crimes, and incarceration by 0.027, 0.018, and 0.021 p.p., respectively, or 11.25%, 18%, and 23.6% of the outcome mean. Teacher effects are thus larger proportionally for more severe CJC outcomes. Substantial effects on both severe crimes and more minor crimes (e.g., traffic tickets) suggest broad influence on criminal behavior as opposed to increased ability to evade detection in crime categories with low clearance rates. 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 future arrests are positively correlated with their effects on the probability of incarceration, as would be expected given that incarceration typically requires a preceding arrest.

3.2 Effects on short-run outcomes

Table 3 presents estimates of the variance-covariance structure of teacher effects on short-run outcomes based on the estimators in Equations 4 and 7. The diagonal entries reflect estimated standard deviations for the outcome in the row/column. The off-diagonals are estimated correlations of effects on the row/column outcomes. The top-left entry, for example, shows that the estimated standard deviation of teacher effects on test scores—combining homeroom, math, and reading teachers—is 0.121. Since test scores are normalized to have a mean of zero and standard deviation of one in the full population of students, this means that a one standard deviation increase in teacher test score quality increases students' scores by 12.1% 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.[16]

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

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

The fifth column shows that the estimated standard deviation of teacher effects on behaviors is 0.125. Recall that the behavioral index is normalized to have a standard deviation of one in our sample, so this estimate also implies that a standard deviation increase in teacher behavioral effects improves behaviors by 12.5% of a standard deviation of the outcome. Interestingly, teacher effects on behaviors are only weakly correlated with teacher effects on test scores. The correlation between behavioral effects and overall test score effects, for example, is 0.056. Similarly, behaviors and study skills effects are only weakly related with a correlation of 0.033. While perhaps surprising, similar results have been found in other contexts. Jackson (2018) and Petek and Pope (2021) find a correlation of 0.15 between teacher value-added on a behavioral index and test scores.[17]

3.3 Predicting teacher effects

Table A.1 examines how teacher effects on short- and long-run outcomes correlate with their demographic and professional observables. In our framework, predicting teacher effects is straightforward since $\bar{Y}_{jt}^k$ or $\bar{Y}_j^k$ can be simply regressed on covariates without further adjustment. We find relatively few strong predictors of teacher impacts. Female, older, better paid, and stronger testing teachers have larger test score effects, for example, but differences are relatively small. The strongest predictor is likely experience, with coefficients that imply an additional 10 years predicts 0.03 higher test score effects. There are few significant predictors of teachers' long-run criminal justice impacts, although female, non-white, and lower-paid teachers tend to have more crime-reducing impacts.

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

3.4 Connecting short- and long-run effects

One way to summarize the relationship between long- and short-run effects is the coefficient from a regression of the former on the latter—e.g., for long-run outcome C and short-run outcome A, $\frac{Cov(\alpha_j^C, \alpha_j^A)}{Var(\alpha_J^A)}$. Using our methods, it is straightforward to obtain a plug-in estimate of this object. Figure 1 reports these estimates for the long- and short-run outcomes studied above. Each coefficient has been re-scaled by the standard deviation of short-run outcome effects so that they can be interpreted as the implied impact on the long-run outcome of exposing a student to a teacher with one standard deviation higher effects on the short-run outcome.

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

Panel (a) presents the results for any CJC, criminal arrests, index crime arrests, and incarceration. Consistent with the small estimated correlations, shifting students to teachers who increase test scores has limited impacts on future arrests. Effects on any CJC and criminal arrests are small, with confidence intervals that include zero. Effects on arrests for index crimes and incarceration are larger but still no more than 0.7 percentage points. Teachers' effects on study skills show similar patterns, with effects statistically indistinguishable from zero. For the purposes of recruiting, retaining, or rewarding teachers who help their students avoid criminal careers, therefore, test score and study skill value-added is not likely to be a particularly useful metric.

Unlike CJC outcomes, we find that teachers who increase test scores do increase students' 12th grade GPA, their likelihood of graduation, and their plans to attend four-year college. The estimated effect of a one standard deviation shift in teacher quality on the latter is roughly 1.3% (roughly 0.6 p.p., similar to the estimated impact

on actual college attendance in Chetty et al. (2014b) of 0.86 p.p.).[18]

By contrast, Panel (b) shows that exposing students to teachers with more positive effects on behaviors has a large impact on future CJC. A one standard deviation shift in behavioral effects is associated with a 2.3% decrease in the likelihood of a criminal arrest and a 3% reduction in the likelihood of being incarcerated between ages 16 to 21. Teacher quality measured through their impacts on these outcomes, therefore, is very relevant for improving students' long-run criminal justice outcomes.[19]

Tables A.2 and A.3 show that we find similar patterns but different magnitudes when using conventional methods that regress students' future CJC outcomes on assigned teachers' value-added estimated in a multi-step EB procedure. For example, assignment to a teacher with a one standard deviation higher test score value-added reduces future criminal arrests by 0.08 p.p. By contrast, our estimate is 0.023 p.p., more than three times smaller and not statistically significant. Conventional estimates suggest assignment to a one standard deviation better behavioral value-added teacher reduces arrests by 0.08 p.p., whereas our estimate implies effects roughly twice as large.

3.5 Multivariate relationships between teacher effects

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

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

[18] It is possible that teachers' effects on future instead of contemporaneous test scores better measure their influence on cognitive skills (Gilraine and Pope, 2021). Figure A.2 shows the long-run impacts of teacher quality as measured by their impacts on test scores and study skills in year $t+1$. These estimates generally show larger impacts, although teacher impacts on behaviors continue to be associated with substantially larger future reductions in CJC.

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

Table 4 presents estimates of $\boldsymbol{\beta}$. Consistent with Figure 1, horse-racing the short-run effects shows that the key predictor of teachers' CJC effects is their effects on behaviors. Test score effects have a negligible relationship with effects on future CJC, while behavioral effects have a much larger one. For 12th grade GPA, graduation, and college attendance, both effects matter independently and substantially.

Given estimates of the total variation in effects on long-run outcomes from Table 2, it is straightforward to calculate the implied R^2 from these regressions. For criminal arrests, the R^2 is 0.042, while for college attendance it is 0.02. Thus, only a small share of the total variance in long-run teacher effects is jointly explained by their 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.

3.6 Validating effects

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

3.6.1 Omitted variables tests of Assumption 1

Interpreting teacher effects as causal under Assumption 1 requires that student unobservables relevant for the outcomes studied are conditionally uncorrelated with teacher assignments. A natural test of this assumption is to assess the sensitivity of teacher effect estimates to the inclusion of proxies for these unobservables (Altonji et al., 2005). To do so, we follow Chetty et al. (2014a) and study the impact of controls excluded from the original model because they are likely unobserved or unused by administrators making teacher assignments. Chetty et al. (2014a) use parental income from merged tax records; we use information on parental education, deeper test score lags, and family fixed effects.

To define the test, consider the "long" regression model that includes the additional

controls W_{it}:

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

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 $\hat{\alpha}_{it}$.[20] 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}, and are therefore potentially useful proxies for student unobservables.

Results: Figure 2 depicts the correlation between estimated teacher effects ($\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$) and predicted outcomes ($\hat{Y}_{it} = W'_{it}\hat{\rho}$) using twice-lagged test scores, parental education, and family fixed effects for twins as omitted variables.[21] 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, implying that the impact of a one standard deviation increase in teacher effects on future arrests ($\sigma^y = 0.027$—see Table 2) may be biased by $0.027 \cdot 0.00135 = 0.000036$ due to omitted variables. Similar results hold for test scores, behaviors, and long-run outcomes, as is shown in Figure A.3 and Tables A.4, and A.5.). Results change little when regressing $\hat{Y}_{it}$ on $\hat{\alpha}_{it}$ in a sample that includes only twins.

Columns 1, 3, and 5 of Table A.4 demonstrate that these omitted variables strongly predict test scores, behavioral measures, and study skills.[22] Table A.5 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

[20] 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.

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

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

of interest. Including them in the teacher effect specification increases the R^2 from 0.089 to 0.107, a 20% increase in the model's explanatory power.

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

3.6.2 Instrumental variable tests of Assumptions 1 and 2

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

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

Assumption 1 implies that $\lambda = 1$ and $\eta_j = 0 \ \forall j$. Assumption 2 implies only that $\lambda = 1$. With an appropriate instrument, it is possible to test whether $\lambda = 1$. 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.[23]

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

[23] 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.

is $Z_{it} = E_{it}\tilde{\alpha}_{sg(i,t)t}$ and is defined analogously at the school level.[24]

We assume that new teacher entry is uncorrelated with student unobservables, or $Cov(Z_{it}, \epsilon_{it}) = 0$. 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.[25] We also assume that the instrument is uncorrelated with teacher-level bias (i.e., $Cov(Z_{it}, \eta_{it}) = 0$) and estimation error in teacher effects (i.e., $Cov(Z_{it}, \xi_{it}) = 0$). Estimating $\tilde{\alpha}_{sgt}$ using all school-grades beside sg bolsters this assumption.

Results: Table 5 reports estimates of λ using teacher switches at the school and school-grade level. Panel (a) shows that for all short-run outcomes we cannot reject $\lambda = 1$. For test scores, for example, the point estimate using teacher switches across school-grades is 1.002 (0.012). Estimates for behavioral measures and study skills are similar, but slightly less precise due to the shorter panel over which they are observed. Estimates for teachers' direct effects on long-run outcomes in Panel (b) are likewise consistent with no bias. In each case, we cannot reject $\lambda = 1$, although λ is less precisely estimated than for short-run outcomes.[26]

The appendix contains several variations on Table 5 that probe the robustness of these results. Table A.6, for example, demonstrates that we also cannot reject unbiased effects when school-grade fixed effects are included, so that only variation in which teachers are assigned to a given school-grade is exploited. Table A.7 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.[27]

[24] Chetty et al. (2014a) and Bacher-Hicks et al. (2014) 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.

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

[26] 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.

[27] Although Rothstein (2017) argues that teacher switches in NC are correlated with student pre-

3.6.3 Are teacher effects actually school effects?

To show that our estimates are not confounded by omitted school effects, we conduct two complementary analyses. The first allows for arbitrary sorting of students within a school, but assumes that any within-school bias is uncorrelated across schools. The second allows for arbitrary sorting of teachers across schools but assumes that assignment of teachers to students within a school satisfies Assumption 1.

To conduct the first exercise, we use teachers who switch schools and examine the relationship between their short- and long-run effects on students in *different* schools. If school effects drove our estimates, we would expect meaningful attenuation, since these schools would likely have different impacts on student outcomes.

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

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

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

In addition to testing for omitted school effects, the estimator defined in Equation 11 significantly weakens our 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 where students who are more likely to excel on standardized tests are assigned to teacher j both in school A and in school B to a similar degree.

Results: Table A.8 reports coefficient estimates from the infeasible regression of long- on short-run effects using this approach. As in the baseline estimates, behavioral effects are strongly related to effects on future CJC, while test score effects are not. The impact of a one standard deviation increase in teacher behavioral effects is similar to our baseline estimate. The standard deviations of teachers' direct effects on long-run outcomes are smaller for some outcomes, but still large. The standard deviation of effects on any criminal arrest, for example, is 1.9 p.p. (vs. 2.7 in the baseline

paredness, our tests only require switches to be conditionally orthogonal to unobserved determinants of outcomes. Table A.7 shows that this holds for CJC.

model).[28]

Our second robustness analysis estimates the variance-covariance of latent teacher effects on all outcomes separately for each school using Equation 4 and computes the teacher-weighted average. These estimates exploit purely within-school variation, washing out any school effects along with any between-school variation in effects.

Results: Table A.9 reports the implied regression coefficients summarizing the relationship between short- and long-run effects using this approach. Estimates are very similar to those in Table 4. Teachers' effects on behaviors strongly predict effects on long-run CJC outcomes, while their effects on test scores are not. 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.

3.6.4 Sensitivity to Assumptions 3 and 4

Table A.10 explores how our key results change when varying the gap between years required for Assumptions 3 and 4 to hold. Adjusting the maximum gap, such that $E[\bar{u}^A_{jt}\bar{u}^C_{jt'}] = 0 \ \forall j, t \neq t', |t - t'| \leq \bar{m}$, is useful for assessing the importance of drift. If teacher effects change substantially over time, setting a smaller $\bar{m}$ should lead to larger estimated variances. Adjusting the minimum gap, such that $E[\bar{u}^A_{jt}\bar{u}^C_{jt'}] = 0 \ \forall j, t \neq t', |t - t'| \geq \underline{m}$, is useful for assessing sensitivity to short-lived runs in student unobservables. The results show a striking degree of stability across $\bar{m}$ and $\underline{m}$. The estimated standard deviations of teacher effects on test scores, behaviors, and criminal justice outcomes change little across assumptions. Our core conclusion that teacher effects on criminal justice outcomes are weakly correlated with test score effects and strongly correlated with behavioral effects is robust to a wide range of choices.

3.6.5 Specification robustness

Even if the controls included in our model are sufficient to account for student observables and unobservables correlated with teacher assignments, our main specification makes several choices over both which specific variables to use and functional form. To explore how sensitive our results are to these modeling choices, we estimate a large number of specifications using 811 different potential sets of controls and, in each case, construct estimates of the impact of a one standard deviation increase in a teacher's

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

test score and behavioral effects on the likelihood of a future criminal arrest.[29] The results reported in Figure A.4 show that our preferred specification is not an outlier. For test scores effects, our preferred estimate is close to the median estimate found when including most potential controls. For behavioral effects, the estimate is among the most conservative possible.

4 Heterogeneous effects

The preceding analysis assumes that teacher effects are the same across students and schools. It is possible, however, that some teachers excel at reaching particular types of students or adjust their teaching priorities based on the classroom environment. 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 $\mu_j(x) = \mu_j + x'\beta_j$ denoting teacher j's effects on students with observables x. Estimating this model allows us to estimate the variance-covariance structure of teachers' effects within and across groups.

The estimation strategy laid out in Section 2.2 makes incorporating heterogeneity in teacher effects simple. For example, to estimate the covariate in teacher effects on test scores between boys and girls, all that is needed is to change the teacher-year average residuals in Equation 7 to be of boys test scores in one year and girls test scores in another year, rather than of test scores and CJC.

We focus on four sets of student characteristics: white vs. non-white students, boys vs. girls, students who are economically disadvantaged vs. not, and student with above vs. below median predicted arrest risk. Table A.11 presents averages of student characteristics and short- and long-run outcomes for these groups. A few disparities are worth noting. White students are meaningfully less likely to have CJC than non-white students, including a seven p.p. (33%) lower likelihood of a criminal arrest and 3.7 p.p. (51%) lower likelihood of being incarcerated. Similarly, girls and students from higher socioeconomic backgrounds have lower CJC rates.

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

Wide variation in the incidence of CJC suggests that teacher effects on CJC may vary considerably across groups. Table A.13 presents estimates of teacher effects on CJC outcomes for different students. Effect variances are often larger for groups with a higher prevalence of CJC, but remain substantial for all student types and for both moderate and more serious contact types. Columns 1 and 2, for example, show that teacher effects are more dispersed for non-white than for white students. Estimates are similar, however, for some groups with large differences in baseline CJC rates. The standard deviation of teacher effects on criminal arrests is 2.72 p.p. for economically disadvantaged students and 2.97 p.p for non-disadvantaged students, despite a nearly 100% difference in average arrest rates.

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

Panels (b) and (c) of Figure 3 show that teachers' effects on long-run outcomes, however, display much weaker correlations for some groups. The correlation of effects on white and non-white students' criminal arrests, for example, is roughly 0.5. As noted earlier, teacher effects on 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.[30]

Finally, we connect these estimates by calculating the implied effects on long-run outcomes of exposing students to teachers with higher student type-specific quality. Figure 4 shows that across various groups, teachers who improve behaviors also reduce the likelihood of future arrests and incarceration. However, as Figure A.5 shows,

[30]Table A.12 shows little evidence of a strong demographic match component to either short- or long-run effects. For test scores, female teachers have marginally higher relative effects on girls' and non-white teachers have marginally higher effects on non-white students. Non-white teachers also tend to reduce crime less for boys, although there is no evidence that teacher race predicts heterogeneous effects on criminal justice outcomes by student race.

across all student types teachers who improve test scores do not reduce CJC. The results highlight the generality of the findings in Figure 1.

5 Implications for teacher retention policies

There has been substantial discussion of how teachers' impacts on student outcomes can be incorporated into retention policies (Rothstein, 2010; Neal, 2011; Chetty et al., 2014a). Ideally, a district would evaluate teachers based on their impacts on students' long-run well being. Doing so is not typically feasible, however, since long-run outcomes are by definition not observed for many years. As a result, in practice teachers are evaluated using their impacts on short-run outcomes such as test scores.

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 illustrate the tradeoffs from emphasizing different measures of teacher quality.[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 do not necessarily reduce the latter. Denote the weighted average of teacher effects on college attendance (μ_j^A) and on future criminal arrests (μ_j^C, signed such that positive values indicate crime reductions) by:

$$\text{Index}_j^{\text{long-run}} = \omega\mu_j^C + (1-\omega)\mu_j^A, \quad \omega \in [0,1] \tag{12}$$

By varying ω, it is straightforward to trace out potential gains from replacing the 5% of teachers with the lowest $\text{Index}_j^{\text{long-run}}$. The rightmost dotted curve in Figure 5 reports the results of this exercise. If long-run effects were directly observed, the district could achieve increases in college attendance of up to 10 p.p. and decreases in criminal arrests of up to 5 p.p. for exposed students. Naturally, increasing one outcome requires reducing effects on the other. Where a district should locate on this frontier depends on their preferences over these long-run goals.

Since teachers' long-run effects are not observed, these estimates represent an upper

[31] To provide a simple connection between our variance estimates and the relevant quantities for these simulations, throughout this section we assume that teacher effects are normally distributed. Appendix D presents more details on the calculations of each policy counterfactual.

bound on what any policy could achieve. In practice, districts must rely on teacher's impacts on short-run outcomes to proxy for teacher quality. The next set of lines in Figure 5 demonstrates the maximum feasible gains from doing so. We construct a weighted average of effects on test scores (μ_j^T), behaviors (μ_j^B), and study skills (μ_j^S) as:

$$\text{Index}_j^{\text{short-run}} = \omega_1 \mu_j^T + \omega_2 \mu_j^B + (1 - \omega_1 - \omega_2)\mu_j^S \tag{13}$$

and examine the impacts of replacing the bottom 5% scoring teachers with the average teacher for different values of $\omega_1 \in [0, 1]$ and $\omega_2 \in [0, 1]$, where $\omega_1 + \omega_2 \leq 1$.

The red dashed line in Figure 5 reports the results of these exercises. Using effects on short-run outcomes, the district could achieve increases of nearly 2 p.p. in college attendance and decreases of no more than about 1 p.p. in criminal arrests for exposed students. Thus, while there are still meaningful potential improvements in long-run outcomes, the frontier lies far to the interior of the infeasible policy.

The green triangle, blue square, and purple circle show the impacts of placing full weight on study skills, behaviors, or test scores, respectively. The score that maximizes impacts on future CJC places 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. This point is slightly inside the frontier, however, demonstrating that even if the district sought to increase college attendance as much as possible they should place at least some weight on behaviors.

In practice, teacher effects on short-run outcomes are also not observed and must be estimated instead. The red dashed-line demonstrates what could be achieved with the best possible estimates, i.e., that coincide with the truth. The costs of estimating scores instead will depend on the information the district has available—how many years teachers are observed, 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 losses 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 et al., 2014a; Gilraine et al., 2021). These results are shown in the solid orange line in Figure 5. Naturally, only a portion of the gains from using true effects on short-run outcomes are achievable when these effects must be estimated. Using our data, the cost implies reductions in

arrests or improvements in college attendance that are roughly half as large.

6 Conclusion

Teachers help students develop a variety of skills necessary to be successful, healthy, and happy adults. The skills needed to excel in one aspect of life, such as the labor market, may differ from those needed in another, such as avoiding entanglement in the criminal justice system. Although prior work demonstrates that teachers who increase students' cognitive skills captured by standardized tests scores increase their college attendance and adult earnings, we find that teachers' test score impacts are orthogonal to students' criminal justice contact as young adults. One of the most common and widespread measures of teacher quality is thus irrelevant for an outcome with life-changing consequences for a large share of the population (Brame et al., 2014).

Instead, teachers who improve proxies for non-cognitive skills such as rates of school discipline and attendance have meaningful impacts on students' future arrest, conviction, and incarceration rates. Our results 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 et al., 2013).

References

Aaronson, D., L. Barrow, and W. Sander (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics 25*, 95–135.

Abaluck, J., M. Caceres Bravo, P. Hull, and A. Starc (2021). Mortality effects and choice across private health insurance plans. *The quarterly journal of economics 136*(3), 1557–1610.

Agostinelli, F. and M. Wiswall (2025). Estimating the technology of children's skill formation. *Journal of Political Economy 133*(3), 846–887.

Altonji, J., T. Elder, and C. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy 113*(1), 151–184.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017, 02). Leveraging Lotteries for School Value-Added: Testing and Estimation. *The Quarterly Journal of Economics 132*(2), 871–919.

Bacher-Hicks, A., S. B. Billings, and D. J. Deming (2019, September). The school to prison pipeline: Long-run impacts of school suspensions on adult crime. Working Paper 26257, National Bureau of Economic Research.

Bacher-Hicks, A., T. J. Kane, and D. O. Staiger (2014). Validating teacher effects estimates using changes in teacher assignments in los angeles. NBER Working Paper No. 20657.

Bacher-Hicks, A. and C. Koedel (2023). Estimation and interpretation of teacher value added in research applications. *Handbook of the Economics of Education 6*, 93–134.

Backes, B., J. Cowan, D. Goldhaber, and R. Theobald (2022). Teachers and students' postsecondary outcomes: Testing the predictive power of test and nontest teacher quality measures. Technical report, CALDER Working Paper.

Bates, M. D., M. Dinerstein, A. C. Johnston, and I. Sorkin (2022). Teacher labor market equilibrium and student achievement. Technical report, National Bureau of Economic Research.

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

Bertrand, M. and J. Pan (2013, January). The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics 5*(1), 32–64.

Biasi, B., C. Fu, and J. Stromme (2021). Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage. Technical report, National Bureau of Economic Research.

Brame, R., S. D. Bushway, R. Paternoster, and M. G. Turner (2014). Demographic patterns of cumulative arrest prevalence by ages 18 and 23. *Crime and Delinquency 60*(3), 471–486.

Chen, J. (2023). Empirical bayes when estimation precision predicts parameters.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9).

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9).

Chetty, R., J. N. Friedman, and J. E. Rockoff (2016, May). Using lagged outcomes to evaluate bias in value-added models. *American Economic Review 106*(5), 393–99.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2017, June). Measuring the impacts of teachers: Reply. *American Economic Review 107*(6), 1685–1717.

Cohen, M. A. and A. R. Piquero (2009). New evidence on the monetary value of saving a high risk youth. *Journal of Quantitative Criminology 25*, 25–49.

Condie, S., L. Lefgren, and D. Sims (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review 40*, 76–92.

Cook, P. J. and S. Kang (2016, January). Birthdays, schooling, and crime: Regression-discontinuity analysis of school performance, delinquency, dropout, and crime initiation. *American Economic Journal: Applied Economics 8*(1), 33–57.

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

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Cunha, F., E. Nielsen, and B. Williams (2021). The econometrics of early childhood human capital and investments. *Annual Review of Economics 13*, 487–513.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review 95*(2), 158–165.

Delgado, W. (2021). Heterogeneous teacher effects, comparative advantage, and match quality. Technical report, Working Paper.

Deming, D. J. (2009, July). Early childhood intervention and life-cycle skill development: Evidence from head start. *American Economic Journal: Applied Economics 1*(3), 111–34.

Deming, D. J. (2011, 10). Better Schools, Less Crime? *The Quarterly Journal of Economics 126*(4), 2063–2115.

Deming, D. J. (2017, 06). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology 92*, 1087–101.

El Karoui, N. and E. Purdom (2018). Can we trust the bootstrap in high-dimensions? the case of linear models. *Journal of Machine Learning Research 19*(5), 1–66.

Gershenson, S., C. M. Hart, J. Hyman, C. A. Lindsay, and N. W. Papageorge (2022). The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy 14*(4), 300–342.

Giacomini, R., S. Lee, and S. Sarpietro (2023). A robust method for microforecasting and estimation of random effects. *arXiv preprint arXiv:2308.01596*.

Gilraine, M., J. Gu, R. McMillan, et al. (2021). A nonparametric method for estimating teacher value-added. Technical report.

Gilraine, M. and N. G. Pope (2021). Making teaching last: Long-run value-added. Technical report, National Bureau of Economic Research.

Gray-Lobe, G., P. A. Pathak, and C. R. Walters (2023). The long-term effects of universal preschool in boston. *The Quarterly Journal of Economics 138*(1), 363–411.

Heckman, J., R. Pinto, and P. Savelyev (2013, October). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–86.

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

Heckman, J. J. and Y. Rubinstein (2001, May). The importance of noncognitive skills: Lessons from the ged testing program. *American Economic Review 91*(2), 145–149.

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

Jackson, C. K. (2014). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics 32*(4), 645–684.

Jackson, C. K. (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. K., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020). School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights 2*(4), 491–508.

Jordan, A., E. Karger, and D. Neal (2024). Early predictors of racial disparities in criminal justice involvement. Technical report, National Bureau of Economic Research.

Kane, T. J. and D. O. Staiger (2008, December). Estimating teacher impacts on student achievement: An experimental evaluation. Working Paper 14607, National Bureau of Economic Research.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.

Kline, P. M., E. K. Rose, and C. R. Walters (2021). Systemic discrimination among large u.s. employers. Technical report, National Bureau of Economic Research.

Krueger, A. B. and L. H. Summers (1988). Efficiency wages and the inter-industry wage structure. *Econometrica: Journal of the Econometric Society*, 259–293.

Kwon, S. (2023). Optimal shrinkage estimation of fixed effects in linear panel data models. *arXiv preprint arXiv:2308.12485*.

Lindqvist, E. and R. Vestman (2011, January). The labor market returns to cognitive and noncognitive ability: Evidence from the swedish enlistment. *American Economic Journal: Applied Economics 3*(1), 101–28.

Lleras, C. (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.

Mulhern, C. and I. Opper (2022). Measuring and summarizing the multiple dimensions of teacher effectiveness.

Neal, D. (2011). The design of performance pay in education. In *Handbook of the Economics of Education*, Volume 4, pp. 495–550.

Papp, J. and M. Mueller-Smith (2021). Benchmarking the criminal justice administrative records system's data infrastructure. Working paper, University of Michigan.

Petek, N. and N. Pope (2021). The multidimensional impact of teachers on students. Technical report, Working Paper.

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

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

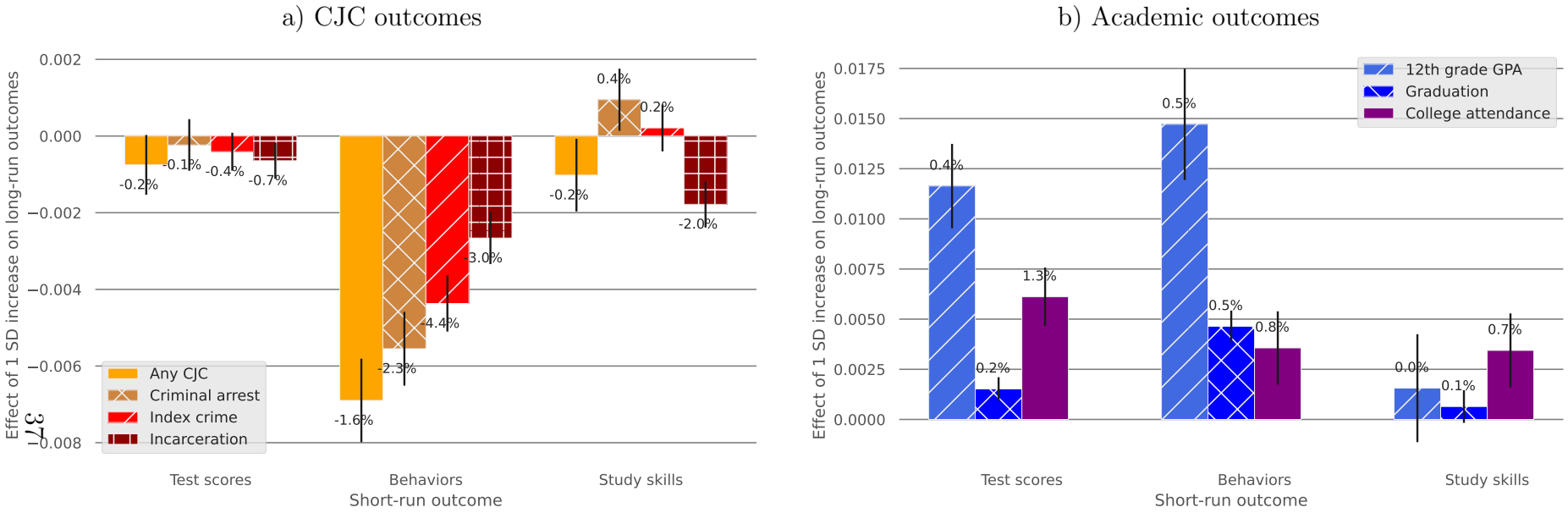
Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics 125*(1), 175–214.

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

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual review of economics 8*(1), 341–377.

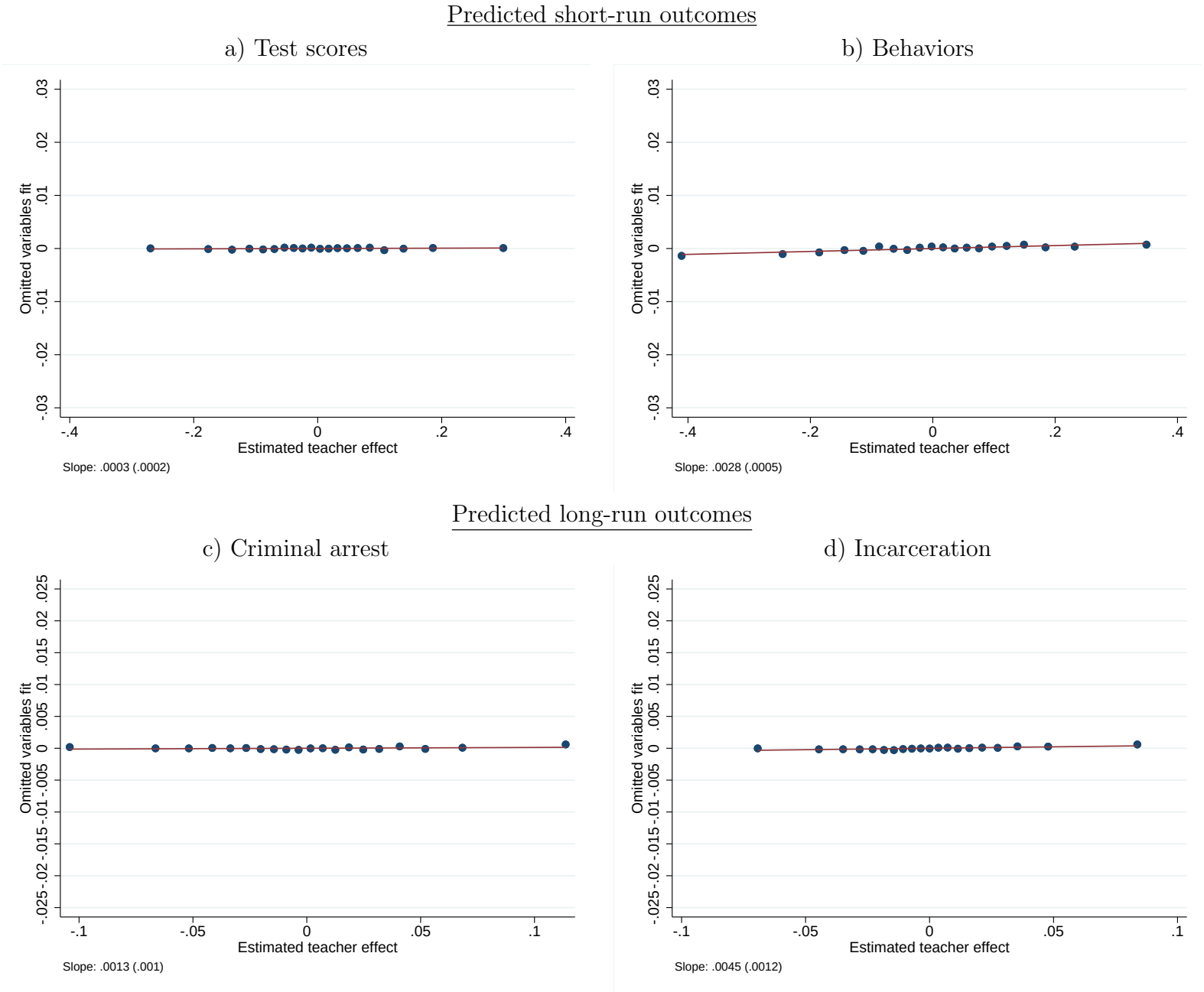
Sorensen, L. C., S. D. Bushway, and E. J. Gifford (2019). Getting tough? the effects of discretionary principal discipline on student outcomes. *Education Finance and Policy*, 1–74.

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



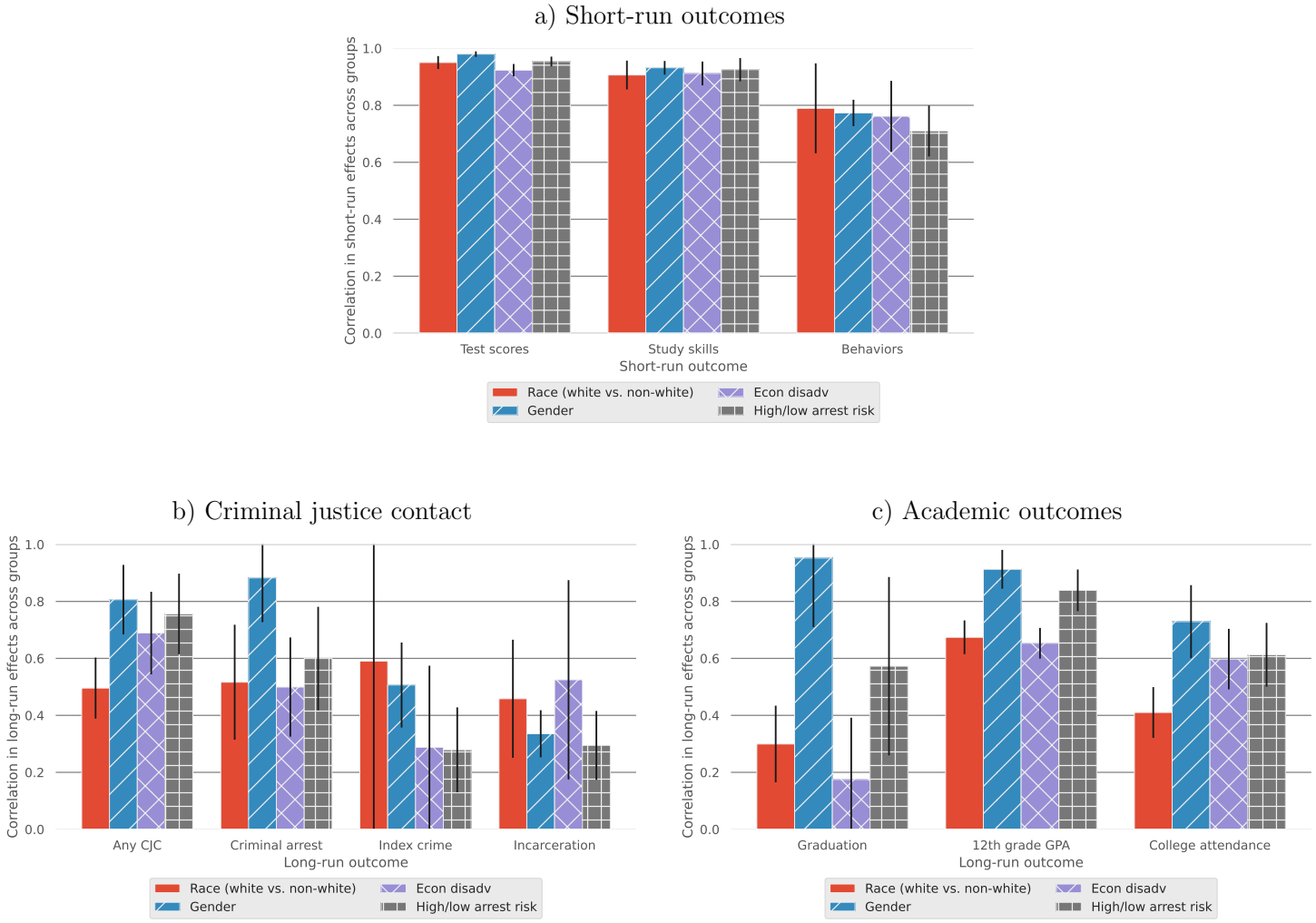
Notes: This figure presents the estimated effect of a 1 standard deviation in teacher quality as measured by short-run outcomes (x-axis) on long-run outcomes implied by estimates of the variance-covariance of teacher effects. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Numbers above/below each bar report effects as a percentage of the outcome mean. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 2.1 and use all available years for each outcome.

Figure 2: Assessing omitted variable bias in teacher effect estimates



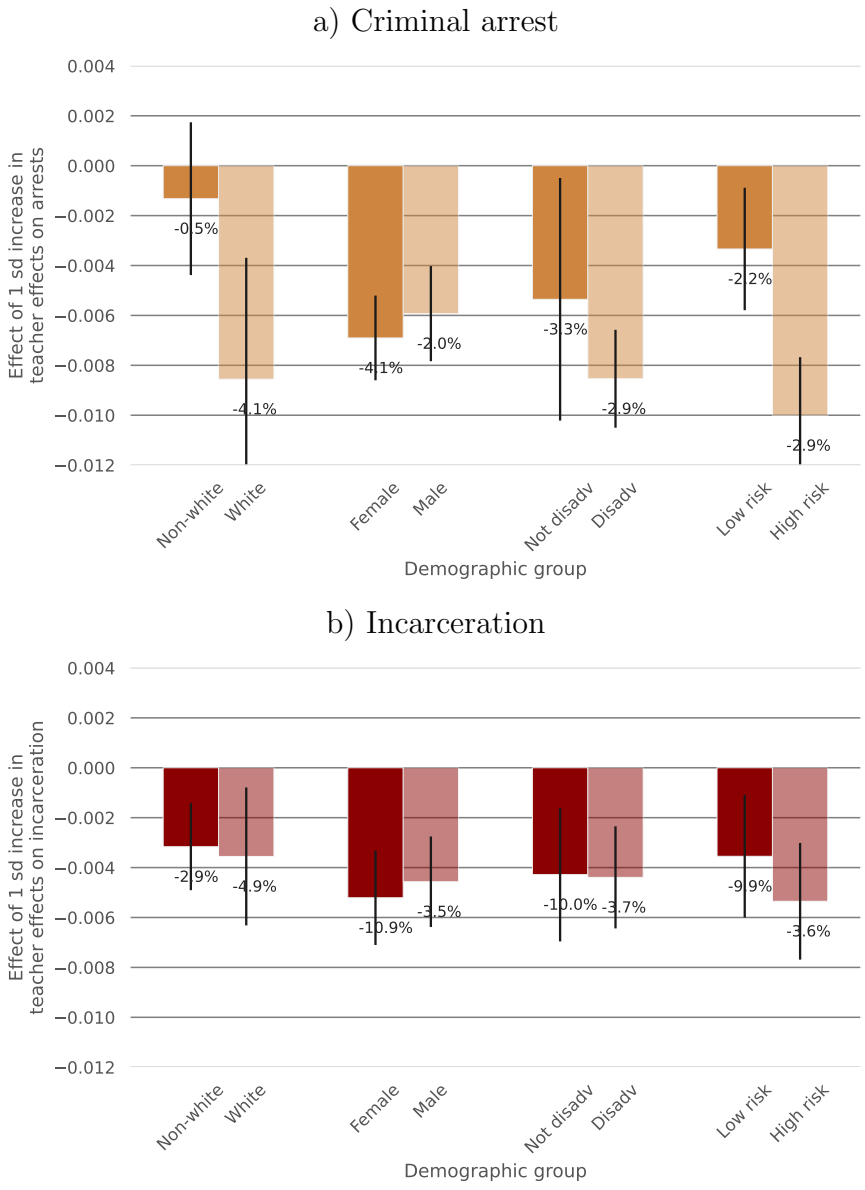
Notes: This figure presents a diagnostic test for whether the estimated teacher effects ($\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$ from Equation 2) are correlated with variables ($W'_{it}\hat{\rho}$ from Equation 9) that predict short- and long-run outcomes but were omitted when estimating the teacher effects. The flat slopes demonstrates that teacher effect estimates are insensitive to the inclusion of these omitted variables. Following Chetty et al. (2014a) 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 2.1 and use all available years for each outcome.

Figure 3: Correlation in teacher effects across groups



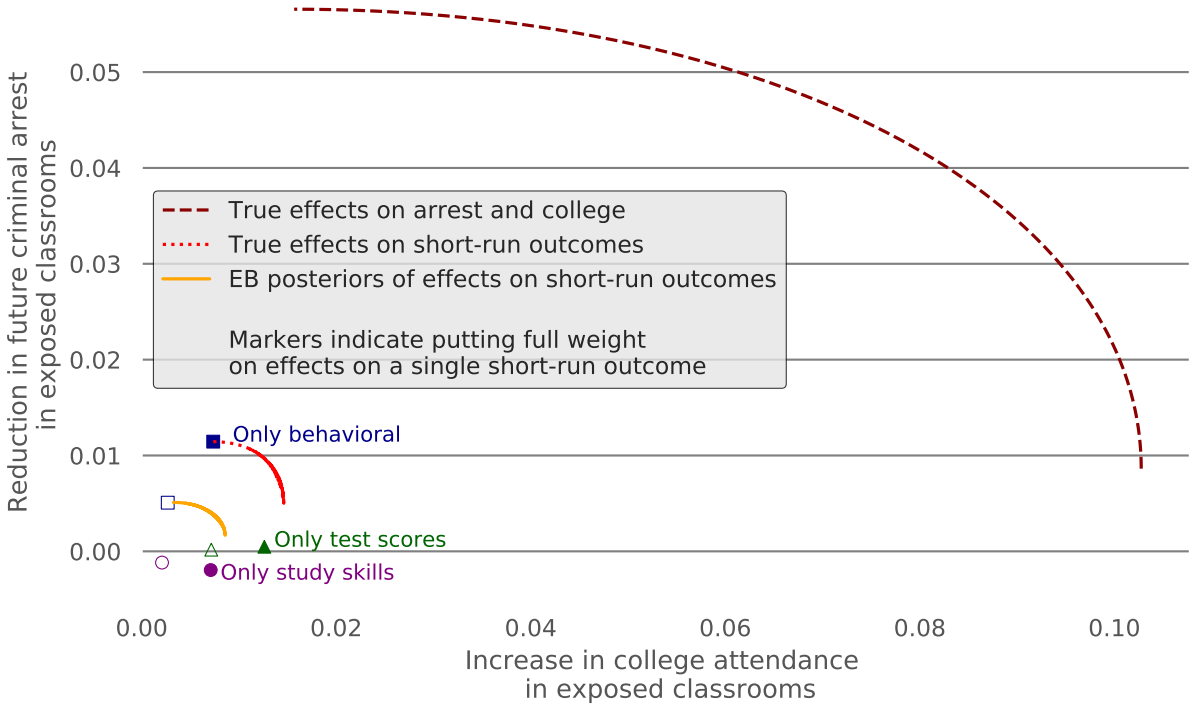
Notes: This figure presents the estimated correlation in teacher effects on short-run outcomes (panel a) and long-run outcomes (panels b and c) across groups of students. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 2.1 and use all available years for each outcome.

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



Notes: This figure presents the estimated effect of a one standard deviation in teacher quality as measured by impacts on students' behaviors on long-run outcomes across groups of students. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 2.1 and use all available years for each outcome.

Figure 5: Effects of teacher removal policies on exposed students



Notes: This figure presents simulations of the impacts of replacing the bottom five percent of teachers with an average teacher on college attendance and future criminal arrests. The rightmost dotted maroon lines in reflect the frontiers achievable if teachers' true long-run effects were directly observed and used to identify which teachers to replace. The dashed red line reflects possibilities if teachers true short-run effects on test scores, behaviors, and study skills were observed and used to select teachers. The leftmost solid line shows possibilities using EB estimates of teacher effects on short-run outcomes instead. The markers indicate gains when putting full weight on a single short-run outcome to select teachers for replacement, with solid markers using true effects and hollow markers using EB posteriors. Teacher effect estimators include the full set of covariates described in Section 2.1 and use all available years for each outcome. All simulations assume teacher effects are jointly normally distributed.

Table 1: Summary statistics

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

Notes: This table presents summary statistics for demographic characteristics, short-run outcomes, and long-run outcomes for the analysis sample, the sample of students for which we observe CJC outcomes, and a sub-sample of students with a criminal arrest between ages 16 to 21. Not all outcomes are observed in all years; summary statistics reflect means and standard deviations for non-missing data only. In each analysis, we use the largest sample possible given when an outcome is observed. See Section 1 for additional details on data construction and outcome coverage by year. Note that the sample of youth with an arrest drops individuals for whom CJC outcomes are unobserved.

Table 2: Direct effects on long-run outcomes

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	College attendance	Graduation
Any CJC	0.035 (0.0007)	0.779 (0.018)	0.513 (0.029)	0.479 (0.025)	-0.055 (0.016)	-0.103 (0.026)	-0.162 (0.025)
Criminal arrest		0.027 (0.0007)	0.816 (0.021)	0.583 (0.023)	-0.150 (0.018)	-0.153 (0.028)	-0.262 (0.029)
Index crime			0.018 (0.0006)	0.507 (0.031)	-0.126 (0.020)	-0.064 (0.032)	-0.352 (0.036)
Incarceration				0.021 (0.0005)	-0.101 (0.016)	-0.175 (0.027)	-0.281 (0.029)
12th grade GPA					0.116 (0.0025)	0.394 (0.032)	0.312 (0.029)
College attendance						0.050 (0.0024)	0.256 (0.041)
Graduation							0.023 (0.0013)

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects on long-run outcomes. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Any CJC refers to any interaction recorded in the criminal justice records between the ages of 16 and 21 inclusive. Criminal arrest excludes non-criminal interactions (e.g., traffic infractions). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' reported plans to attend a four-year college reported after graduation. Graduation is an indicator for graduating high school. Teacher effect estimators include the full set of covariates described in Section 2.1 and using all available years for each outcome.

Table 3: Teacher effects on short-run outcomes

	Test scores	Math scores	Reading scores	Study skills	Behaviors
Test scores	0.121 (0.0006)	0.909 (0.0020)	0.810 (0.0035)	0.317 (0.0071)	0.056 (0.0089)
Math scores		0.134 (0.0005)	0.675 (0.0057)	0.279 (0.0067)	0.047 (0.0089)
Reading scores			0.073 (0.0006)	0.337 (0.0090)	0.071 (0.0109)
Study skills				0.183 (0.0018)	0.033 (0.0085)
Behaviors					0.125 (0.0017)

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects on short-run outcomes. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Teacher effect estimators include the full set of covariates described in Section 2.1 and using all available years for each outcome.

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

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	-0.001 (0.004)	-0.003 (0.003)	-0.003 (0.002)	0.000 (0.002)	0.097 (0.010)	0.010 (0.003)	0.045 (0.007)
Behaviors	-0.059 (0.005)	-0.048 (0.004)	-0.038 (0.003)	-0.023 (0.003)	0.124 (0.013)	0.039 (0.004)	0.029 (0.008)
Study skills	-0.004 (0.003)	0.007 (0.002)	0.002 (0.002)	-0.009 (0.002)	-0.014 (0.008)	0.001 (0.002)	0.009 (0.006)
$sd(\mu_j^y)$	0.035 (0.001)	0.027 (0.001)	0.018 (0.001)	0.021 (0.001)	0.116 (0.003)	0.023 (0.001)	0.050 (0.002)
R^2	0.039	0.042	0.059	0.023	0.025	0.041	0.020

Notes: This table presents the coefficients from a regression of long-run outcomes on short-run teacher effects implied by variance-covariance matrix of short- and long-run teachers effects. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. The final two rows report estimated standard deviations of teacher effects on the long-run outcome and the R^2 from the regression. See also the notes of Table 3.

Table 5: Instrumental variables tests for forecast unbiased teacher effects

(a) Short-run outcomes

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

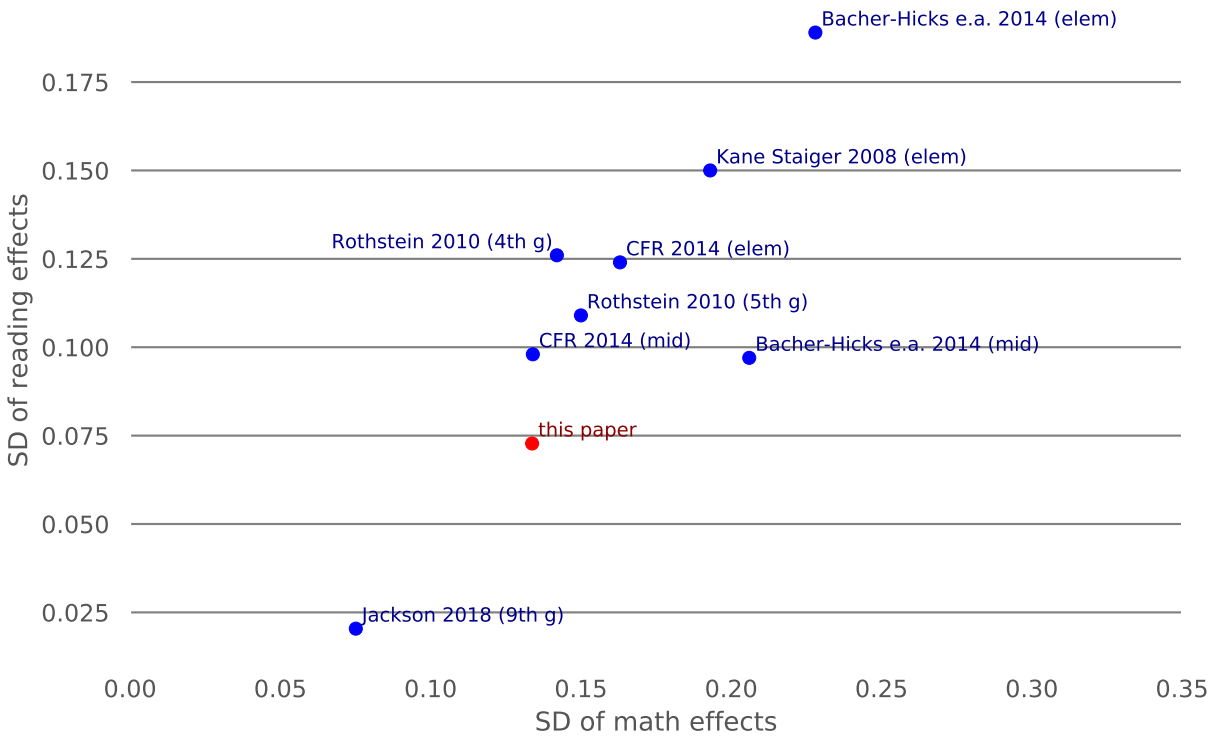
(b) Long-run outcomes

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

Notes: This table presents instrumental variable tests for bias in estimated teacher effects on short- and long-run outcomes, where an estimate of 1 implies forecast unbiased estimates. Design controls include the full set of covariates described in Section 2.1 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.

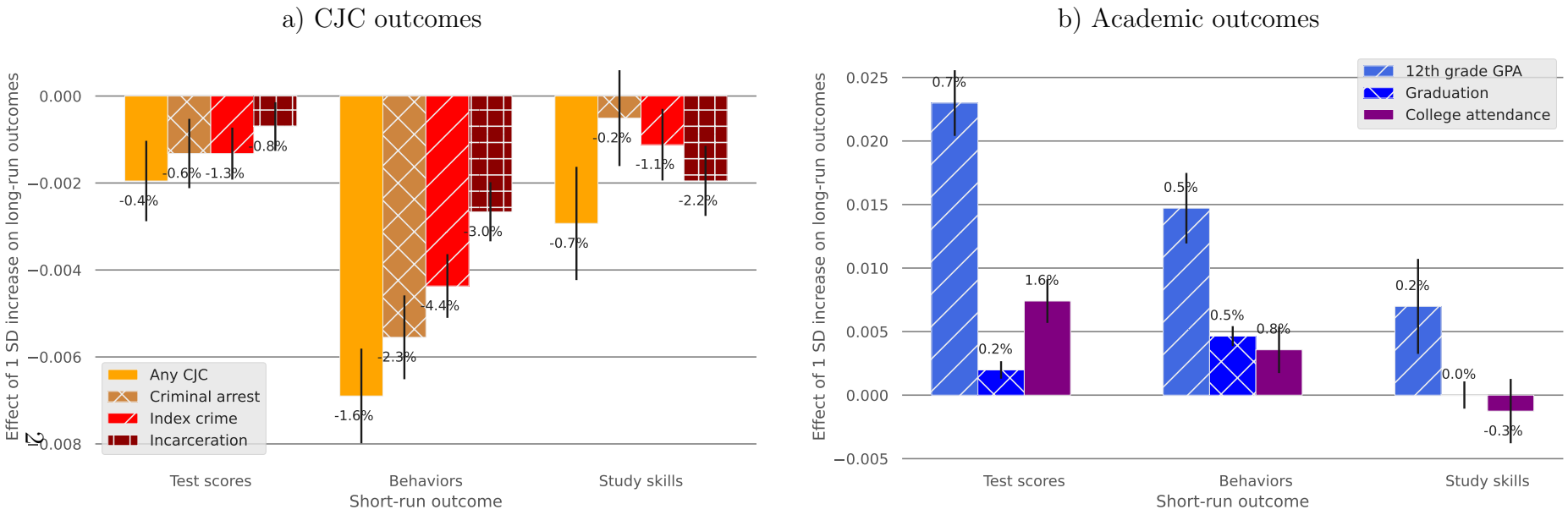
A Additional figures and tables

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: Effects of "long-run" teacher quality on outcomes

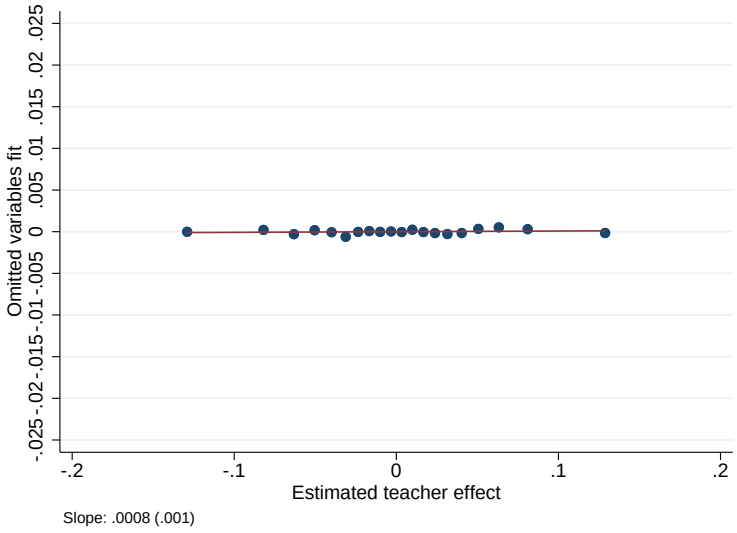


Notes: This figure presents the estimated effect of a 1 standard deviation in teacher quality as measured by short-run outcomes (x-axis) on long-run outcomes implied by estimates of the variance-covariance of teacher effects. Unlike in Figure 1, teachers effects on test scores and study skills are measured using outcomes in $t+1$ to capture their "long-run" impacts (Gilraine and Pope, 2021). Behavioral outcomes were already measured at $t+1$ in our primary analysis, but their effects are repeated here for comparison. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Numbers above/below each bar report effects as a percentage of the outcome mean. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 2.1 and use all available years for each outcome.

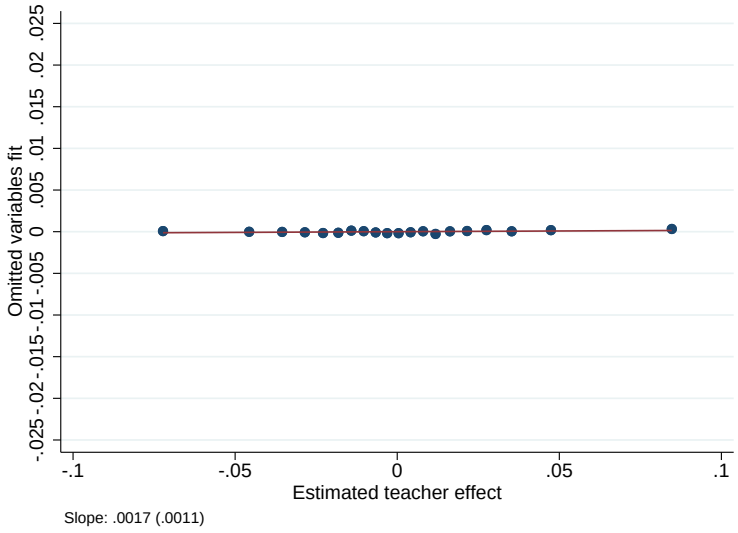
Figure A.3: Assessing omitted variable bias in additional long-run outcomes

Predicted long-run CJC outcomes

a) Any CJC

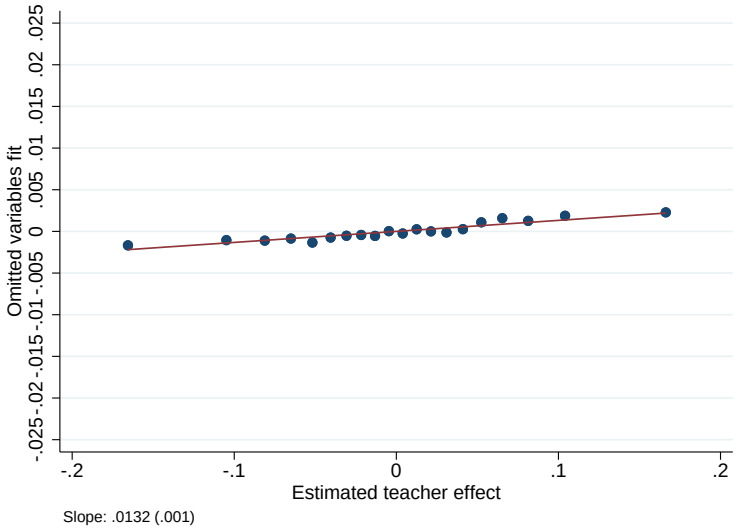


b) Arrest for index crime

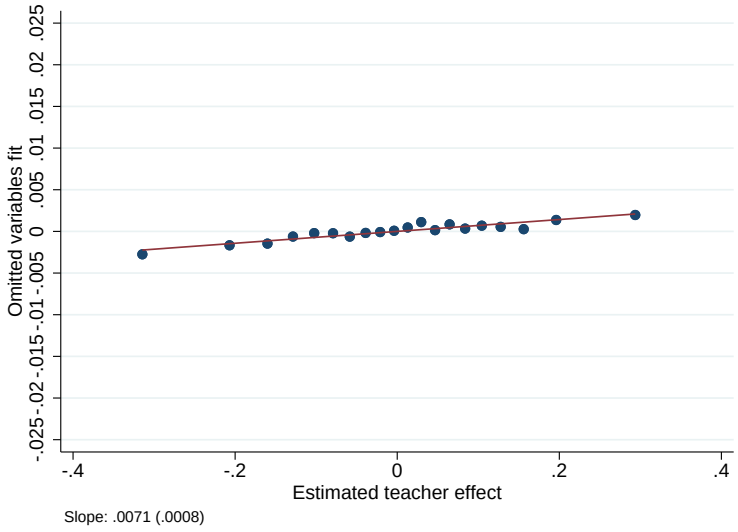


Predicted long-run academic outcomes

c) College bound

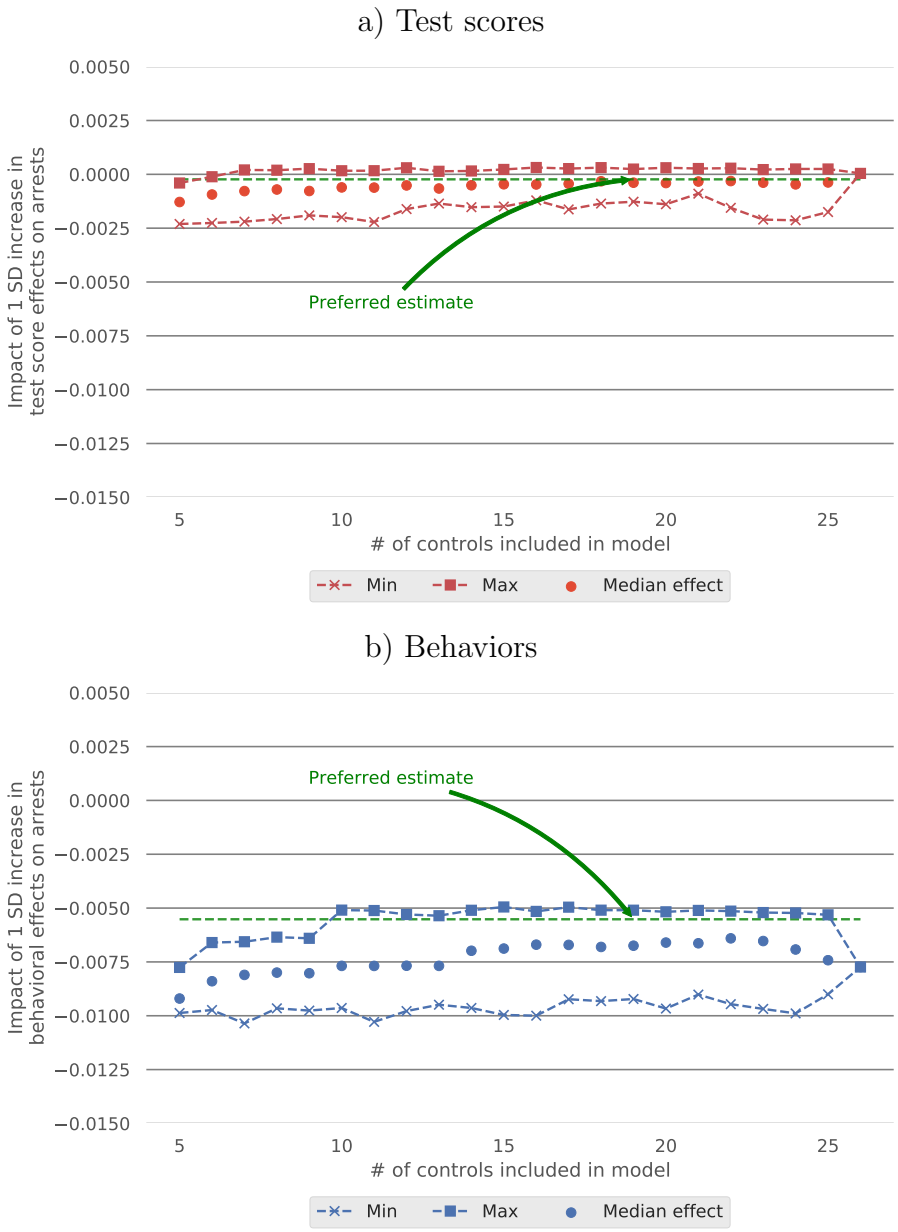


d) 12th grade GPA



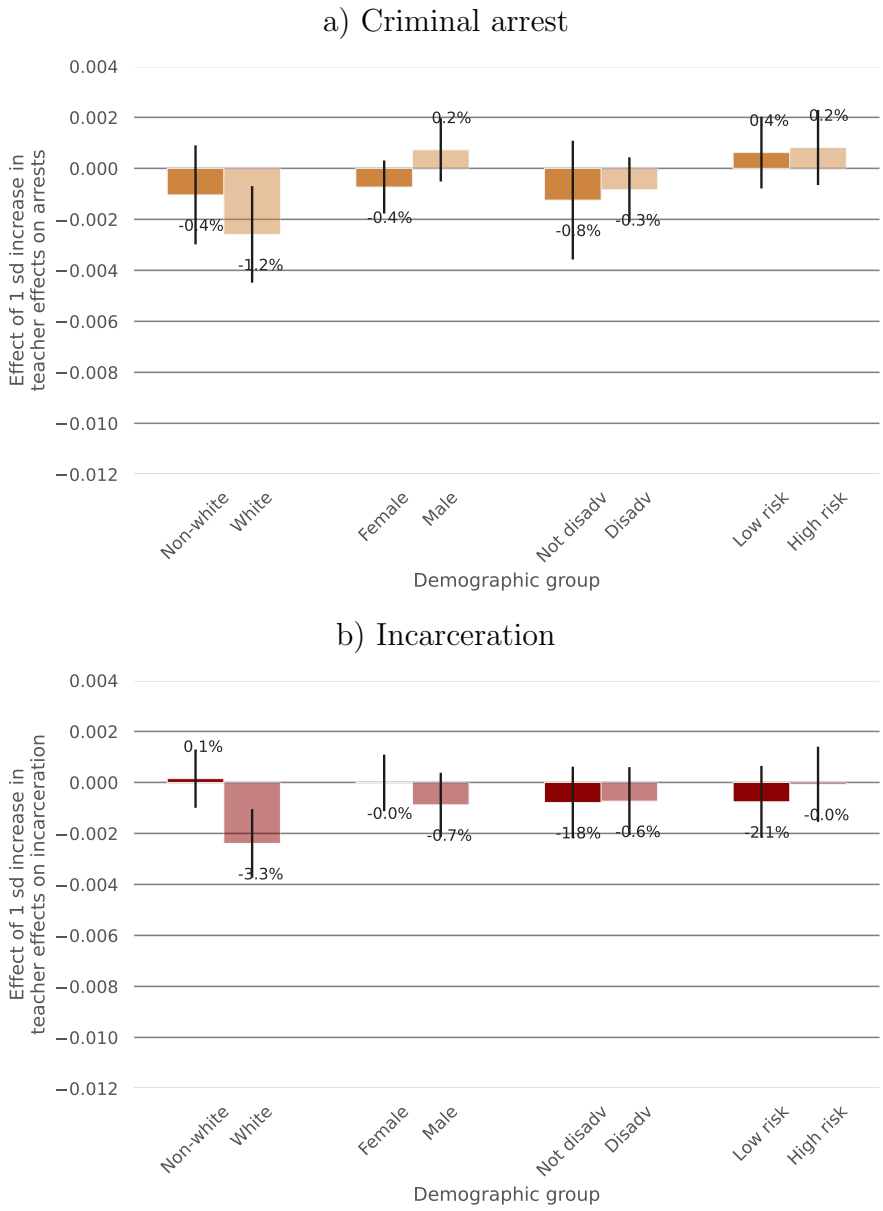
Notes: This figure presents a diagnostic test for whether the estimated teacher effects ($\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$ from Equation 2) are correlated with predictions based on omitted variables ($W'_{it}\hat{\rho}$ from Equation 9) that are predictable of the short- and long-run outcomes but have not been used when estimating the teacher effects. Following Chetty et al. (2014a) 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 2.1 and use all available years for each outcome.

Figure A.4: Specification sensitivity of teacher quality impacts on arrests



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

Figure A.5: Heterogeneous impacts of exposure to teachers who improve test scores



Notes: This figure presents the estimated effect of a one standard deviation in teacher quality as measured by impacts on students' test scores on long-run outcomes across groups of students. The error bars are 95% confidence intervals based on analytic standard errors estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. Teacher effect estimators include the full set of covariates described in Section 2.1 and use all available years for each outcome.

Table A.1: Predictors of teacher effects

	(1) Test scores	(2) Study skills	(3) Behaviors	(4) Criminal arrest	(5) Index crime	(6) Incarceration
Female	0.0251***	0.0632***	0.00684	-0.00348***	-0.00170*	-0.00195**
	(0.00233)	(0.00447)	(0.00364)	(0.00102)	(0.000735)	(0.000752)
Non-white	-0.0126***	-0.0271***	-0.00561	-0.00116	-0.000332	-0.00506***
	(0.00247)	(0.00502)	(0.00407)	(0.00116)	(0.000869)	(0.000866)
Age	0.000171*	0.00128***	-0.000351**	0.0000268	-0.0000210	0.0000287
	(0.0000784)	(0.000169)	(0.000125)	(0.0000351)	(0.0000251)	(0.0000254)
Prior experience	0.00318***	0.00335***	0.000900*	0.0000238	0.00000751	0.000119
	(0.000223)	(0.000745)	(0.000367)	(0.000123)	(0.0000876)	(0.0000865)
Masters or higher	0.00362*	0.0130**	0.00115	-0.00115	0.0000686	0.000516
	(0.00171)	(0.00404)	(0.00261)	(0.000778)	(0.000552)	(0.000561)
Average test score	0.00696***	0.0153***	-0.00362*	0.000544	0.000327	-0.000383
	(0.00105)	(0.00224)	(0.00168)	(0.000465)	(0.000327)	(0.000338)
Pay (standardized)	0.00807***	0.00781***	-0.00424**	0.000688	0.000339	-0.00134***
	(0.000802)	(0.00163)	(0.00149)	(0.000391)	(0.000303)	(0.000301)
Constant	-0.0393***	-0.130***	0.00470	0.00289	0.00284*	0.000940
	(0.00356)	(0.00726)	(0.00567)	(0.00160)	(0.00114)	(0.00114)
N teacher-years	196344	89947	87684	82228	82228	82228
N teachers	35044	24604	24695	24232	24232	24232

Notes: This table presents demographic and professional predictors of teacher effects on multiple outcomes. Each column is a single regression of $\bar{Y}_{jt}^{k}$, where k is the outcome listed in the column header, on several teacher characteristics for all teacher-years in which both the outcome and the teacher characteristic are available. Age and prior experience are measured in years. Average test score is the teacher-year average of standardized (within test-type and year) scores on Praxis licensing exams. Pay is gross total pay standardized to have mean zero and standard deviation one within each year.

Table A.2: Regression based estimates of teacher test score 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
Test score VA	-0.0108	-0.00795	-0.00591	-0.00553	0.0781	0.0104	0.0368
	(0.00398)	(0.00332)	(0.00228)	(0.00232)	(0.00907)	(0.00220)	(0.00478)
Design controls	✓	✓	✓	✓	✓	✓	✓
1SD effect	-.0012	-.0008	-.0006	-.0006	.0083	.0011	.0039
R2	0.0480	0.0754	0.0596	0.0663	0.542	0.106	0.253
Observations	4159500	4159500	4159500	4159500	3429388	4623602	3205422

Notes: This table presents regressions of teacher test score value added calculated using the method in Chetty et al. (2014a) 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. Standard errors clustered at the student level are reported in parentheses.

Table A.3: 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	-0.0517	-0.0453	-0.0354	-0.0233	0.179	0.0259	0.0361
	(0.00639)	(0.00541)	(0.00376)	(0.00365)	(0.0130)	(0.00346)	(0.00753)
Design controls	✓	✓	✓	✓	✓	✓	✓
1SD effect	-.0042	-.0036	-.0028	-.0019	.0142	.002	.0029
R2	0.0481	0.0751	0.0596	0.0651	0.543	0.106	0.256
Observations	3700227	3700227	3700227	3700227	2969019	4001899	2850488

Notes: See the notes of Table A.2.

Table A.4: Omitted variables bias tests for short-run teacher effects

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

Notes: This table presents tests for omitted variable bias in estimated teacher effects on short-run outcomes. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. The odd columns regress the outcome listed in the sub-header on the excluded covariates, teacher dummies, and the full set of design controls described in Section 2.1. The even columns regress predicted outcomes based on the excluded covariates on estimated teacher effects $\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$. Education variables refer to students' reported parental education, with an indicator for missing parental education data serving as the omitted category. Lag 2 math and reading refer to twice-lagged standardized test scores, with indicators for missing twice-lag scores included but not reported. Twin-effects include fixed effects for all twin pairs and an indicator for non-twin interacted with year. Results change little when regressing $\hat{Y}_{it}$ on $\hat{\alpha}_{it}$ in the sample of twins only. Original R^2 refers to 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.5: 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	0.00300		0.0177		0.0198		0.0270	
	(0.00113)		(0.000954)		(0.000694)		(0.000642)	
High school only	-0.0000113		0.0184		0.0147		0.0173	
	(0.000866)		(0.000730)		(0.000531)		(0.000492)	
Some college	-0.00307		-0.0170		-0.0150		-0.0175	
	(0.000913)		(0.000770)		(0.000560)		(0.000518)	
BA or more	-0.0196		-0.0178		-0.00873		-0.00615	
	(0.000781)		(0.000658)		(0.000479)		(0.000443)	
Lag 2 math	0.0107		0.00268		-0.000369		0.000680	
	(0.000562)		(0.000474)		(0.000345)		(0.000319)	
Lag 2 reading	-0.00953		-0.00776		-0.00544		-0.00499	
	(0.000526)		(0.000443)		(0.000322)		(0.000298)	
Teacher effect		0.000794		0.00128		0.00169		0.00452
		(0.00101)		(0.00103)		(0.00115)		(0.00119)
Observations	4145532	4145532	4145532	4145532	4145532	4145532	4145532	4145532
R2	0.0811	0.718	0.107	0.702	0.0914	0.790	0.101	0.737
Original R2	.062		.0888		.0726		.0807	
Design controls	✓	✓	✓	✓	✓	✓	✓	✓
Twin FE	✓		✓		✓		✓	

Notes: This table presents tests for omitted variable bias in estimated teacher effects on long-run outcomes. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. The odd columns regress the outcome listed in the sub-header on the excluded covariates, teacher dummies, and the full set of design controls described in Section 2.1. The even columns regress predicted outcomes based on the excluded covariates on estimated teacher effects $\hat{\alpha}_{it} = \sum_j \hat{\alpha}_j D_{ijt}$. Education variables refer to students' reported parental education, with an indicator for missing parental education data serving as the omitted category. Lag 2 math and reading refer to twice-lagged standardized test scores, with indicators for missing twice-lag scores included but not reported. Twin-effects include fixed effects for all twin pairs and an indicator for non-twin interacted with year. Results change little when regressing $\hat{Y}_{it}$ on $\hat{\alpha}_{it}$ in the sample of twins only. Original R^2 refers to 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: 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
$\hat{\alpha}_j$	1.002	1.052	1.255	1.681	1.083	1.126
	(0.0135)	(0.0175)	(0.270)	(1.101)	(0.0641)	(0.0825)
Observations	9779708	9779708	5422682	5422682	3404657	3404657
R^2	0.723	0.721	0.209	0.200	0.135	0.132
Design controls	✓	✓	✓	✓	✓	✓
School-grade FE	✓	✓	✓	✓	✓	✓
First stage F	47960	30550	364	24	2691	1709
P-value for $H_0 : \lambda = 1$	.875	.003	.346	.536	.198	.127

Notes: This table presents instrumental variable tests for bias in estimated teacher effects on short-run outcomes. Design controls include the full set of covariates described in Section 2.1 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.7: Instrumental variables bias tests for teacher-effects on criminal arrest

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

Notes: This table presents instrumental variable tests for bias in estimated teacher effects on future criminal arrests. 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. See also the notes to Table A.6.

Table A.8: Implied regression of long-run effects on short-run effects using only teachers who move across schools

	Any CJC	Criminal arrest	Index crime	Incarceration	12th grade GPA	Graduation	College attendance
Test scores	0.009 (0.009)	-0.011 (0.007)	-0.002 (0.005)	0.006 (0.005)	-0.054 (0.028)	-0.005 (0.007)	0.025 (0.017)
Behaviors	-0.051 (0.029)	-0.033 (0.025)	-0.043 (0.021)	-0.024 (0.018)	0.277 (0.093)	0.000 (0.020)	0.103 (0.046)
Study skills	-0.014 (0.008)	0.021 (0.007)	0.000 (0.005)	-0.018 (0.005)	0.089 (0.024)	0.027 (0.007)	0.035 (0.016)
$sd(\mu_j^y)$	0.023 (0.003)	0.019 (0.003)	0.009 (0.004)	0.014 (0.002)	0.072 (0.009)	0.012 (0.007)	0.033 (0.009)
R^2	0.026	0.045	0.108	0.038	0.084	0.095	0.077

Notes: This table presents the coefficients from a regression of long-run outcomes on short-run teacher effects implied by variance-covariance matrix of short- and long-run teachers effects. The estimates are based variance-covariance estimated from leaving-one-out teacher-school pairs rather than the leave-one-out teacher-year estimates reported in Table 4. Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Test scores refers to the combined outcome of the first principal component of math and reading scores for homeroom teachers, math scores for math teachers, and reading scores for reading teachers, respectively. Behaviors refers to the first principal component of an indicator for any discipline in $t+1$, total days absent in year $t+1$, and an indicator for grade repetition, all standardized within year and grade. Study skills refers to the first principal component of standardized within year and grade reported time spent on homework, watching TV, and reading. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). 12th grade GPA is a six-point-scale GPA for the student's first appearance in 12th grade. College attendance is an indicator for students' plans to attend a four-year college reported in 12th grade. The final two rows report estimate standard deviations of teacher effects on the long-run outcome and the R^2 from the regression. Teacher effect estimators include the full set of covariates described in Section 2.1 and using all available years for each outcome.

Table A.9: Implied regressions using within-school variation

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

Notes: This table presents the coefficients from a regression of long-run outcomes on short-run teacher effects implied by the variance-covariance matrix of short- and long-run teachers effects using only *within* school variation in teacher effects. Estimates are based on the primary model in which teacher effects are constant across schools and time. See also the notes of Table A.8

Table A.10: Sensitivity to runs and drift

a) Test scores, [Study skills], {Behaviors} Standard deviations

min m	max m = 1	2	3	∞
	0.131	0.128	0.126	0.121
1	[0.196]	[0.191]	[0.188]	[0.183]
	{0.143}	{0.132}	{0.129}	{0.125}
		0.120	0.117	0.112
2		[0.183]	[0.179]	[0.173]
		{0.108}	{0.108}	{0.103}
			0.113	0.106
3			[0.175]	[0.167]
			{0.107}	{0.098}

b) Any CJC, [Crim arrest], {Index crime} Standard deviations

min m	max m = 1	2	3	∞
	0.037	0.036	0.035	0.035
1	[0.029]	[0.029]	[0.028]	[0.027]
	{0.020}	{0.019}	{0.019}	{0.018}
		0.033	0.032	0.031
2		[0.026]	[0.025]	[0.024]
		{0.017}	{0.015}	{0.015}
			0.030	0.029
3			[0.024]	[0.023]
			{0.013}	{0.013}

c) Any CJC, [Crim arrest], {Index crime} Correlation with test score effects

min m	max m = 1	2	3	∞
	−0.019	−0.019	−0.017	−0.022
1	[−0.012]	[−0.011]	[−0.007]	[−0.008]
	{−0.023}	{−0.027}	{−0.028}	{−0.023}
		−0.013	−0.005	−0.019
2		[−0.024]	[−0.008]	[−0.013]
		{−0.041}	{−0.034}	{−0.027}
			−0.023	−0.035
3			[−0.023]	[−0.025]
			{−0.039}	{−0.026}

d) Any CJC, [Crim arrest], {Index crime} Correlation with behavior effects

min m	max m = 1	2	3	∞
	−0.177	−0.190	−0.202	−0.199
1	[−0.212]	[−0.227]	[−0.220]	[−0.202]
	{−0.266}	{−0.303}	{−0.284}	{−0.242}
		−0.208	−0.246	−0.262
2		[−0.245]	[−0.252]	[−0.258]
		{−0.342}	{−0.315}	{−0.288}
			−0.307	−0.326
3			[−0.305]	[−0.312]
			{−0.340}	{−0.303}

Notes: This table illustrates the sensitivity of our estimates to variations on Assumption 3 and 4. Our baseline estimates assume that $E[\bar{u}^A_{jt}\bar{u}^C_{jt'}] = 0\ \forall j, t \neq t'$. Each cell in this table shows estimates assuming it holds only for min $m \leq |t - t'| \leq$ max m. The minimum m corresponds to the rows, while the maximum m is listed in each column. our baseline estimates correspond to min $m = 1$ and max $m = \infty$. Panel a presents estimated standard deviations of effects on test scores, study skills, and behaviors. Panel b presents analogous estimates for any CJC, criminal arrests, and index crimes. Panels c and d present correlations between effects on these criminal justice outcomes and test score effects (panel c) and for behavioral effects (panel d).

Table A.11: Summary statistics of four different sub-groups comparisons across race, sex, socioeconomic status, and predicted risk of arrest

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

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

Table A.12: Predictors of teachers' heterogeneous demographic effects

a) Boys - Girls effects

	(1) Test scores	(2) Study skills	(3) Behaviors	(4) Criminal arrest	(5) Index crime	(6) Incarceration
Female	-0.0135***	0.0192***	-0.00963*	-0.00179	-0.00210	-0.00623***
	(0.00160)	(0.00528)	(0.00485)	(0.00198)	(0.00151)	(0.00147)
Non-white	-0.00515**	0.0215***	-0.0107*	0.00467*	0.0224***	0.0211***
	(0.00163)	(0.00543)	(0.00539)	(0.00210)	(0.00171)	(0.00160)
Constant	-0.0116***	-0.287***	-0.186***	0.126***	0.0399***	0.0849***
	(0.00149)	(0.00490)	(0.00455)	(0.00184)	(0.00141)	(0.00137)
N teachers	39670	26856	28439	27214	27214	27214

b) White - Non-white effects

	(1) Test scores	(2) Study skills	(3) Behaviors	(4) Criminal arrest	(5) Index crime	(6) Incarceration
Female	0.0105***	0.00196	-0.00729	0.000734	0.00185	-0.000448
	(0.00234)	(0.00742)	(0.00724)	(0.00296)	(0.00226)	(0.00213)
Non-white	-0.0105***	0.0220**	0.0252**	-0.00623*	0.000117	-0.000323
	(0.00287)	(0.00830)	(0.00887)	(0.00317)	(0.00243)	(0.00231)
Constant	-0.00733***	0.00729	-0.110***	0.0173***	0.00511*	0.00186
	(0.00218)	(0.00691)	(0.00681)	(0.00277)	(0.00212)	(0.00200)
N teachers	39292	26167	27763	26477	26477	26477

Notes: This table presents demographic predictors of teachers' relative effects on boys vs. girls (panel a) and white vs. non-white students (panel b). The outcome in each regression is the difference in the teacher-level average of $\bar{Y}_{jt}^k$ across groups, with positive quantities indicating larger effects on boys (in panel a) or white students (in panel b). Each regression includes all teachers where it is possible to estimate $\bar{Y}_{jt}^k$ for each sub-group and where teacher demographics are available.

Table A.13: Heterogeneity in teacher effects on CJC

	Race		Sex		Econ. disadvantaged		Arrest risk	
	(1) White	(2) Non-white	(3) Boys	(4) Girls	(5) Yes	(6) No	(7) High	(8) Low
Any CJC	0.0391	0.0370	0.0412	0.0364	0.0364	0.0367	0.0443	0.0436
	(0.00195)	(0.00184)	(0.00259)	(0.00161)	(0.00131)	(0.00134)	(0.00225)	(0.00178)
Criminal arrest	0.0247	0.0358	0.0298	0.0311	0.0272	0.0297	0.0342	0.0263
	(0.00274)	(0.00186)	(0.00358)	(0.00176)	(0.00114)	(0.00142)	(0.00308)	(0.00309)
Index crime	0.0177	0.0248	0.0147	0.0239	0.0212	0.0209	0.0276	0.0113
	(0.00290)	(0.00203)	(0.00625)	(0.00181)	(0.000967)	(0.00127)	(0.00339)	(0.00694)
Incarceration	0.0172	0.0308	0.0156	0.0270	0.0211	0.0303	0.0320	0.0201
	(0.00267)	(0.00142)	(0.00512)	(0.00136)	(0.000696)	(0.00102)	(0.00285)	(0.00380)
Observations								

Standard errors in parentheses

Notes: This table presents estimates of the standard deviations of teacher effects on future CJC across various sub-populations: white vs. non-white (Columns 1 and 2), boys vs. girls (Columns 3 and 4), economically disadvantaged vs. not (Columns 5 and 6), and students with high vs. low predicted risk of a future arrest (Columns 7 and 8). Analytic standard errors displayed in parentheses are estimated using the procedure described in Appendix C. Criminal arrest refers to any criminal interaction with the justice system between ages 16 and 21 (i.e., excluding traffic tickets and non-criminal violations). Traffic citations includes only non-criminal traffic violations. Index crimes includes arrests for Uniform Crime Reporting index crimes: aggravated assault, forcible rape, murder, robbery, arson, burglary, larceny/theft, and motor vehicle theft. Incarceration refers to any incarceration sentence in local jails or state prisons. Teacher effect estimators include the full set of covariates described in Section 2.1 and using all available years for each outcome.

B What does the covariance of EB posteriors estimate?

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

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

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

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

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

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

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

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

For simplicity, assume that each teacher is assigned n students with outcomes gener-

ated by this process. The average outcome of students assigned to teacher j is:

$$\bar{Y}_j^k = \alpha_j^k + \bar{u}_j^k$$

where $\bar{u}_j^k = \frac{\sum_{i=1}^n \epsilon_i^k 1(j(i)=j)}{n}$. Due to the normal-normal structure of the model, the univariate EB posterior for teacher j's effect on outcome k, or $E[\alpha_j|\bar{Y}_j]$, is simply $\lambda_{EB}^k \bar{Y}_j^k$, where

$$\lambda_{EB}^k = \frac{(\sigma_\alpha^k)^2}{(\sigma_\alpha^k)^2 + Var(\bar{u}_j^k)}$$

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

$$Cov(\lambda_{EB}^A \bar{Y}_j^A, \lambda_{EB}^C \bar{Y}_j^C) = \lambda_{EB}^A \lambda_{EB}^C \left[\overbrace{Cov(\alpha_j^A, \alpha_j^C)}^{\text{Cov. in teacher effects}} + \underbrace{Cov\left(\bar{u}_j^A, \bar{u}_j^C\right)}_{=Cov(\epsilon_i^A, \epsilon_i^C)/n}{}^{\text{Cov. in sampling error}} \right]$$

The direction of bias depends on two terms: attenuation due to $\lambda_{EB}^A \lambda_{EB}^C$, which falls between zero and one, and correlated sampling error $Cov\left(\bar{u}_j^A, \bar{u}_j^C\right)$. When the latter is zero, the covariance of EB posteriors is attenuated toward zero (and the correlation is biased upwards) and could in principle be corrected by undoing multiplication by $\lambda_{EB}^A \lambda_{EB}^C$, although this is rarely done. Since in general $Cov\left(\bar{u}_j^A, \bar{u}_j^C\right)$ (i.e., $Cov\left(\epsilon_i^A, \epsilon_i^C\right)$) can take any sign, the overall bias is unclear in general settings.

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

We use the estimated variance-covariance of teacher effects and classroom-level sampling error to construct the implied covariance (and correlation) in EB posteriors from the normal-normal model described above for different values of n and examine how it

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

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

Notes: This table presents estimated standard deviations (diagonal elements) and correlations (off-diagonal elements) of teacher effects and classroom-level sampling error (i.e., $\bar{u}_j^k$) for key outcomes. The former is reported without brackets, while the latter is reported in square brackets.

relates to the true covariances (and correlations) of latent effects. The EB estimates use a multivariate model that constructs posterior means as:

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

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

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

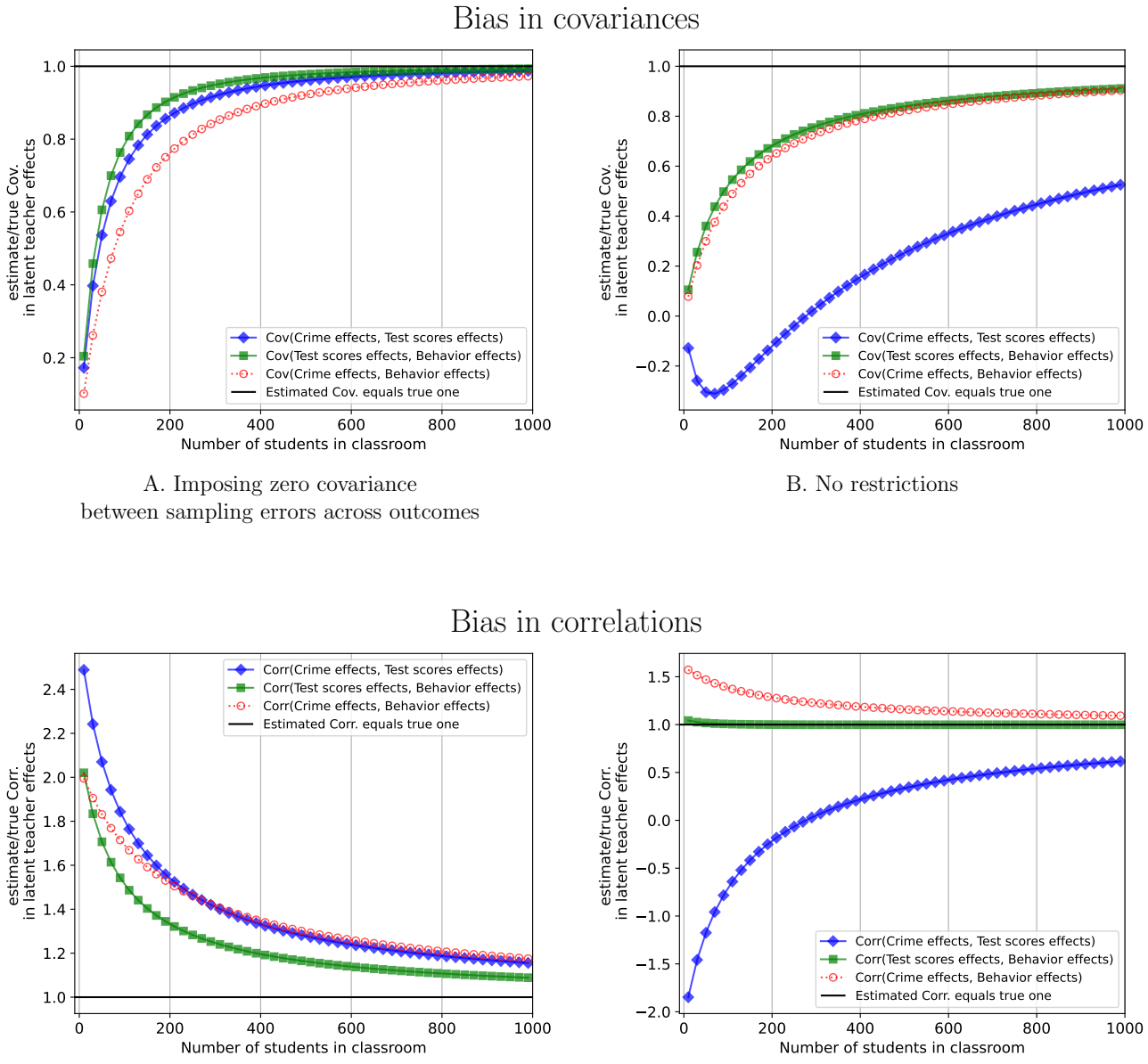
In Panel A, we impose that there is no correlated classroom-level sampling error across outcomes, so that $Cov\left(\bar{u}_j^k, \bar{u}_j^{k'}\right)$ (i.e., $Cov(\epsilon_i^k, \epsilon_i^{k'}) = 0$) for any two outcomes k and k'. As we would expect, the bias is decreasing with the number of students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

In Panel B, we re-introduce the correlations in classroom-level sampling error reported in Table B.1. The results are consistent across the outcomes considered. The biases from correlation in classroom-level sampling error across outcomes and the shrinkage factors are large and persistent. Importantly, for the covariance of test scores and criminal arrest effects the bias is large even with 1,000 students per teacher, implying

very accurate estimates of each teacher's effects.

Panels C and D report the bias of estimates of correlations rather than covariances and are analogous to Panels A and B. The results are broadly similar and show persistent and meaningful bias even with a large number of students per teacher. For the correlation between effects on behaviors and test scores, the biases from correlated classroom-level sampling error across outcomes and the shrinkage factors happen to cancel each other out. The former biases the estimate up, while the latter pulls the estimate towards zero. However, for the other correlations the biases do not cancel out. Importantly, for the correlation in test scores and criminal arrest the bias is meaningful even with 1,000 students per teacher, implying very accurate estimates of each teacher's effects.

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



A. Imposing zero covariance between sampling errors across outcomes

B. No restrictions

C. Imposing zero covariance between sampling errors across outcomes

D. No restrictions

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

C Inference on variance components

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

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

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

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

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

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

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

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

which can be written as:

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

where $(\sigma_i^A)^2 = \sigma_i^{AA}$.

C.1 Plug-in estimator of standard errors

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

$$Var(\hat{\lambda}_i^{ml}) = \sum_{k\neq i} (C_{ik}^{ml})^2 \sigma_k^{ll}$$

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

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

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

D Policy simulation details

This appendix includes technical details for the implementation of the simulations discussed in Section 5. These simulations examine the implications for a given long-run outcome of replacing the bottom 5% of teachers, according to a given measure of quality, with an average teacher for exposed students. In Section 5, we discuss ranking teachers based on three different options: (i) an index based on teachers' true direct effect on long-run outcomes, (ii) an index using teachers' true effects on short-run outcomes, and (iii) an index using Empirical Bayes estimates of teacher effects on short-run outcomes. In all cases, we assume that all short- and long-run teacher effects are jointly normally distributed, allowing us to characterize the full distribution of teacher quality using our variance estimates.

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

In this case, the calculations are straightforward. We are interested in the impact of replacing the bottom 5% of teachers according to the quality index in Equation 12 with the average teacher. Thus, when estimating the effect of such a policy on teachers' effect on college attendance, for example, the estimand of interest is:

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

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

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

In this case, the calculations are also straightforward. We are interested in the impact of replacing the bottom 5% of teachers according to the quality index in Equation 13 with average teachers. Thus, when estimating the effect of such a policy on teachers' effect on college attendance, for example, the estimand of interest is:

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

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

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

In this case, the calculations require a few steps. Recall that we are interested in the impact of replacing the bottom 5% of teachers with average teachers according to an Empirical Bayes estimate of the quality index in Equation 13.

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

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

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

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

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

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

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

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

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

the predicted change in Y due to the replacement policy:

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

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

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

and note that:

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

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