Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1170

Disparate Teacher Effects, Comparative Advantage, and Match Quality

William Delgado
University of Chicago

Does student-teacher match quality exist? While prior research documents disparities in teachers' impacts across student types, it has not distinguished between sorting and causal effects as the drivers of these disparities. I develop a flexible disparate value-added model (DVA) and introduce a novel measure of teacher quality–revealed comparative advantage (CA)–that captures the degree to which teachers affect student outcome gaps. Leveraging a quasi-experimental teacher turnover design, I show that the CA measure accurately predicts teachers' disparate impacts: a teacher with a 1 standard deviation in black CA increases black students' test scores by 1 standard deviation, with no effect on non-black students' test scores. This methodological contribution offers a framework to study match effects, with implications for policy efficiency and equity.

VERSION: April 2025

Suggested citation: Delgado, William. (2025). Disparate Teacher Effects, Comparative Advantage, and Match Quality. (EdWorkingPaper: 25 -1170). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/nr10-5s67

Disparate Teacher Effects, Comparative Advantage, and Match Quality

William Delgado *

February 2025

Abstract

Does student-teacher match quality exist? While prior research documents disparities in teachers' impacts across student types, it has not distinguished between sorting and causal effects as the drivers of these disparities. I develop a flexible disparate value-added model (DVA) and introduce a novel measure of teacher quality—revealed comparative advantage (CA)—that captures the degree to which teachers affect student outcome gaps. Leveraging a quasi-experimental teacher turnover design, I show that the CA measure accurately predicts teachers' disparate impacts: a teacher with a 1 standard deviation in black CA increases black students' test scores by 1 standard deviation, with no effect on non-black students' test scores. This methodological contribution offers a framework to study match effects, with implications for policy efficiency and equity.

Keywords: teacher quality, value-added, comparative advantage, match quality, achievement gaps

JEL codes: H75, I21, I24, J24, J45

---

*Delgado: UChicago Consortium on School Research, Crown Family School of Social Work, Policy, and Practice, University of Chicago (email: wdelgado@uchicago.edu). I thank Elaine Allensworth, John Q. Easton, and the staff at the UChicago Consortium on School Research for helping me use Chicago Public Schools administrative data and better understand the policy context. I am extremely grateful to Seth Zimmerman, Dan Black, Damon Jones, Stephen Raudenbush, and Susan E. Mayer for helpful comments and numerous conversations. I am also grateful to Josh Goodman, Marcus Winters, Andrew Bacher-Hicks, Olivia Chi, Kirsten Slungaard Mumma, Alexis Orellana and colleagues and students at Boston University Wheelock College of Education and Human Development for comments and conversations. I also thank Lauren Sartain, Martha Olney, Peter Q. Blair, John Friedman, Robert Kaestner, Doug Staiger, Jesse Rothstein, Conrad Miller, Raj Chetty, Douglas Miller, Michael Dinerstein, Ellora Derenoncourt, Jack Mountjoy, Valerie Michelman, Joshua Goodman, Paula Worthington, Oren Danieli, Deborah Weiss, Randall P. Ellis and anonymous referees, and many others for their feedback and comments. I thank seminar participants at the Boston University Empirical Microeconomic Workshop, UChicago Workshop on Education, Economics of Racism Working Group, Cornell University, Ohio University, ASSA, AEFP, SOLE, LACEA-LAMES, and MEA. This research received no specific grants. Any errors are my own.

1 Introduction

Racial, gender, and income achievement gaps are a persistent, stubborn issue in the U.S. and globally (Fryer Jr, 2011; Reardon et al., 2019a,b; Michelmore and Dynarski, 2017; Chmielewski, 2019). Research shows that these achievement gaps contribute to an array of disparities in adult outcomes, including the racial gap in earnings and the gender gap in college enrollment (Neal and Johnson, 1996; Fryer Jr, 2011; Aucejo and James, 2021). Teachers play an important role in improving student achievement and long-term outcomes (Chetty et al., 2014b; Jackson, 2018; Gershenson et al., 2022). However, existing disparities in teacher impacts across student demographics remain poorly understood. While prior studies have documented differences in teacher effectiveness across student subgroups, it is unclear whether these differences arise from causal effects or systematic sorting of students to teachers. Understanding this distinction is crucial for designing equitable and efficient education policies and closing opportunity gaps.

This paper develops a methodological framework to test for student-teacher match quality and proposes a novel measure of teacher quality–revealed comparative advantage (CA). CA captures a teacher's differential impact on one student subgroup relative to another. I ask (i) To what extent do teacher impacts on student outcomes vary across subgroups? If so, (ii) how can incorporating these effects improve policy efficiency and equity? The primary contribution of this work lies in extending the value-added (VA) framework by incorporating flexibility to estimate heterogeneous teacher impacts while addressing key identification challenges. The empirical section serves as an application of the proposed method, showing its utility in quantifying teacher impacts across student subgroups and informing policy decisions.

I develop a disparate value-added (DVA) model that allows teacher effects to vary flexibly across dimensions of student heterogeneity. Unlike traditional VA models that assume homogeneous impacts, the DVA framework captures both within- and across-teacher variability in teacher effects. The proposed CA measure captures the within-teacher variation and is validated using a teacher-switching quasi-experiment, which leverages changes in teaching staff across schools and grades.

To illustrate the quasi-experimental strategy, suppose a mixed-race school has one classroom in 4th grade whose teacher is teacher A. Also suppose that in the following year, this teacher is replaced by teacher B, who is otherwise similar except that her black CA quality is *predicted* to be 1σ higher. If teacher CA is an unbiased predictor of disparate teacher effects, then one should observe a 1σ increase in black students' test scores and no change in non-black students' test scores when teacher A is replaced by teacher B. This quasi-experimental design ensures that observed changes in student outcomes are attributable to teacher effects rather than systematic sorting. The results show that CA is an unbiased predictor of teachers' impacts, providing robust evidence of student-teacher match quality.

To illustrate the policy implications of these findings, I simulate teacher-to-classroom assignments and teacher accountability policies that incorporate CA. For example, matching teachers with high CA for specific subgroups to classrooms with higher proportions of those students generates substantial efficiency gains in student outcomes. These simulations also reveal important trade-offs between equity and efficiency, highlighting the complexity of optimizing teacher assignments under resource constraints.

This paper contributes to three main strands of literature: teacher value-added models, student-teacher match effects, and the broader field of labor economics and workers' comparative advantage. By proposing a novel value-added model to semi-parametrically estimate teacher impact disparities and identify individual teachers' comparative advantage, this paper bridges gaps in prior research and highlights critical implications for policy and practice. The next Section relates the paper's contribution to the literature. Section 3 outlines the conceptual framework and develops the DVA model. Section 4 develops the quasi-experimental test for student-teacher match quality. Section 5 describes the data and Section 6 applies the DVA framework to study teacher impacts on racial achievement gaps. Section 7 presents quasi-experimental results and validation of the teacher CA measure. Section 8 discusses counterfactual policy simulations and explores potential sources of heterogeneity in teacher CA. The final section concludes.

2 Related literature

Teacher value-added models. The literature on teacher value-added models has extensively documented disparities in teacher impacts across different student subgroups, defined by race/ethnicity, gender, socioeconomic status, prior achievement levels, and other characteristics. These disparities suggest that teachers can have differential impacts on specific groups of students. However, methodologies and conclusions in this area vary significantly.

For instance, Loeb et al. (2014) estimate teacher value-added separately for English learners (ELs) and non-ELs using a value-added model with teacher fixed effects. By applying a Bayesian shrinkage procedure to mitigate noise in individual estimates, they find high but imperfect correlations between VA for ELs and non-ELs (0.89 for math, 0.80 for reading). Their findings suggest significant overlap in teacher effectiveness across groups, yet highlight some disparities that could inform targeted student-teacher assignments. Master et al. (2016) find that prior experience teaching and specific training experiences focused on ELLs improve teacher effectiveness with ELLs.

Similarly, Condie et al. (2014) employ a hierarchical linear model (HLM) to estimate teacher VA for high- and low-achieving students. With correlations ranging from 0.97 (math) to 0.8–0.9 (reading), they find that matching teachers to students based on comparative advantage could increase reading achievement by 0.02 standard deviations on average. Their counterfactual exercises show the potential for welfare gains, particularly through optimal student-to-teacher assignments. Fox (2016) extends this analysis by examining teacher VA across multiple student subgroups, including gender, poverty status, and race/ethnicity using HLM. While subgroup correlations generally exceed 0.9, his findings suggest minimal disparities in teacher effects across these dimensions. These results, however, are based on the assumption of random teacher-student assignments and may underestimate heterogeneity in teacher impacts.

A more recent study by Barrios-Fernández and Riudavets-Barcons (2024) explores teacher effectiveness across gender subgroups using a value-added model with drift separately fit to female and male students. They find significant disparities in gender-specific teacher impacts, with correlations of 0.73 (math), 0.69 (reading), and as low as 0.35–0.36 for educational attainment index, which combines high school completion, registration for the university admission exam, enrollment in higher education, enrollment in university, and enrollment in a selective university. Their counterfactual simulations suggest that eliminating within-teacher disparities could substantially reduce gender gaps in math by 67 percent, Spanish test scores gap by 15 percent, and educational attainment by 10 percent.

Student-teacher match effects. A key contribution of this paper is to advance the literature on student-teacher match effects by proposing a quasi-experimental design to test for their existence. Prior research has primarily focused on demographic match effects, such as race and gender congruence. For example, studies have found that students perform better when matched with teachers of the same race or gender (Dee, 2007; Egalite et al., 2015; Gershenson et al., 2022). These effects extend to long-term outcomes, including college enrollment and aspirations (Egalite and Kisida, 2018) and teacher expectations (Gershenson et al., 2016).

However, this paper moves beyond demographic matching to estimate individual teachers' comparative advantage across multidimensional student characteristics. This approach allows for a more nuanced understanding of how specific teachers impact specific subgroups. For example, Ahn et al. (2024) use a flexible parametric model to estimate teacher effects as a linear function of student characteristics (e.g., race, gender, prior achievement, poverty status). They find substantial within-teacher heterogeneity, with the variance of the match component being 0.08 standard deviations.[1] These findings align with the semi-parametric approach proposed in this paper, which further tests whether observed disparities arise from sorting or causal effects.

Validation of novel measures and policy implications. The third major contribution of this paper is to validate a novel measure of teacher quality, referred to as comparative advantage. This builds on existing work that seeks to validate value-added measures as estimates of causal teacher effects (Kane and Staiger, 2008; Chetty et al., 2014a, e.g.,). By explicitly modeling disparities in teacher impacts and linking them to quasi-experimental design, this paper provides validation for using comparative advantage in policy decisions.

Recent studies have explored the welfare consequences of considering teacher heterogeneity in labor market equilibrium models. For instance, Biasi et al. (2021) find that allowing districts to optimize teacher assignments based on comparative advantage could improve total teacher contributions by 2.7 percent but exacerbate achievement gaps. Similarly, Bates et al. (2024) simulate counterfactuals to show how addressing teacher-preferences and principal-hiring practices could simultaneously close the baseline achievement gap by one-fourteenth and increase average student achievement.

This paper complements these studies by quantifying the potential equity and efficiency gains

[1] Other examples include Lockwood and McCaffrey (2009), who estimate heterogeneity in teacher effects across students of different achievement levels using a HLM-analogue model. They find strong evidence of interaction effects, with match effects accounting for up to 10 percent of the variance in impacts; however, they conclude that ignoring these interactions is not likely to introduce appreciable bias in estimated teacher effects. Graham et al. (2020) interacts student baseline achievement with teaching practices, measured by Danielson's Framework for Teaching (FFT) instrument, and find strong complementaries. Students with high baseline test scores score higher when matched with high-FFT teachers.

of considering teacher comparative advantage in policies such as teacher-to-classroom reallocations and professional development programs. For example, this paper's simulations show how optimal teacher assignments can reduce racial achievement gaps while maintaining or improving overall student outcomes—an important policy goal for addressing educational opportunity gaps.

3 Conceptual framework

3.1 Theories on sources of disparate teacher effects

What causes disparities in a teacher's impact on student outcomes across student subgroups? One possible explanation is that the teacher differentially allocates inputs across student subgroups, and this unequal allocation produces unequal outcomes. For example, a teacher with conscious or unconscious bias against minority students may spend less individualized time with minority students, encourage them less often, give them different types of advice, grade them more strictly, or punish them more harshly compared to non-minority students (e.g., Lindsay and Hart, 2017; see Redding, 2019, for a review). A teacher without cultural competency may relate differently to students from different cultures (Irvine, 1989).

Another explanation for the disparities in teacher effects is that the teacher may homogeneously allocate inputs across students but student subgroups respond differentially to the inputs. Role model effects and ethnic studies curriculum are examples where the teacher teaches the same materials to her students but some student subgroups benefit more, perhaps because the teacher inspires them for sharing similar background characteristics or the course material motivates them for being related to their background (e.g., Dee and Penner, 2017; Dee, 2004). Distinguishing between the two explanations may have different policy implications. If disparities in teacher effects are due to teacher behavior, then providing professional development may be an avenue to reduce disparities in student outcomes. If it is due to role model effects or other student behavioral responses, then policies such as increasing teacher diversity may be a better avenue. Whichever the source, a teacher's disparate effects are revealed by the observed disparities in her students' outcomes, and the DVA model presented below aims to estimate these disparate effects.

3.2 Disparate value-added model

Let i index students and t years. Assume that each student i is assigned to a classroom $c = c(i,t)$. Assume also that each teacher teaches one classroom per year, and let $j = j(c(i,t))$ denote student i's teacher in year t. Each student belongs to one of K mutually exclusive and collectively exhaustive types, denoted by $k(i) \in \{1, ..., K\}$. Each teacher is allowed to affect these student types differentially, either because she differentially allocates inputs across student subgroups or her inputs have different productivities or both. These effects are captured by μ_{jkt}, which represents teacher j's VA in year t for students of type k. The main difference between this DVA model and a typical VA model is that a teacher has multiple effects rather than a single one that affects all students homogeneously. I scale each teacher's VA within student types so that the average teacher's effects are $\mu_{j1t} = ... = \mu_{jKt} = 0$, and a one unit increase in each teacher effect corresponds to a 1 standard deviation increase in student test scores.

The outcome of student i in year t, A^*_{it}, is given by

$$A^*_{it} = X'_{it}\beta_{k(i)} + \nu_{ik(i)t}, \tag{1}$$

where

$$\nu_{ikt} = \mu_{jkt} + \theta_{ck} + \varepsilon_{ikt}. \tag{2}$$

Equation 1 states that student i's outcome can be decomposed into an observable part, X_{it}, and an unobservable part, ν_{ikt}. The observable part includes student characteristics, such as baseline test scores, gender, race, and ethnicity as well as classroom and school characteristics. The unobserved part has three components: a student-type-specific teacher effect, μ_{jkt}; a student-type-specific classroom shock, θ_{ck}; and a student-level idiosyncratic shock that may depend on student type, ε_{ikt}. Student-type-specific classroom shocks affect students from the same type equally. An example of such shocks is police violence against black teenagers, which differentially affects the performance of minority and non-minority students in schools close to those events (Ang, 2021). Note that the coefficient β_k varies by student type, allowing for differential effects of observable characteristics on student outcomes that could result from systemic discrimination or other constraints differentially affecting student types.

Teacher quality is not observed, but it can be inferred through its impact on student outcomes. If students were randomly assigned to teachers, the average increase in (unexplained) test scores for students of type k in teacher j's classroom would estimate teacher j's type-k-specific VA. However, due to non-random sorting of students to teachers, X_{it} and ε_{ikt} may be correlated with μ_{jkt}, and thus estimates of teacher quality using observational data may be biased. The quasi-experimental strategy aims to address this concern.

I make the following identifying assumption to estimate teachers' student-type-specific effects:

Assumption 1 (Joint stationarity) *Student-type-specific teacher effects, student-type-specific classroom shocks, and individual-level shocks follow a stationary process:*

$$\mathbb{E}[\mu_{jkt}|k,t] = \mathbb{E}[\theta_{ck}|k,t] = \mathbb{E}[\varepsilon_{ik(i)t}|k,t] = 0$$

$$Cov(\mu_{jkt}, \mu_{jm,t+s}) = \sigma_{\mu_k\mu_m,s}$$

$$Cov(\theta_{ck}, \theta_{cm}) = \sigma_{\theta_k\theta_m}$$

$$Cov(\varepsilon_{ik(i)t}, \varepsilon_{ik(i),t+s}) = \sigma_{\varepsilon_k,s}$$

for all t, $s \geq 0$, and $k, m \in \{1, ..., K\}$.

Assumption 1 has several implications.[2] First, the mean of each student-type-specific teacher effect is constant across years. Second, the autocorrelation of teacher quality ($\sigma_{\mu_k,s} \equiv \sigma_{\mu_k\mu_k,s}$), cross-correlation of teacher quality ($\sigma_{\mu_k\mu_m,s}$), and autocorrelation of individual-level shocks ($\sigma_{\varepsilon_k,s}$) only depend on the amount of time elapsed. Third, the cross-correlation of teacher quality across two different years is symmetric, irrespective of which effect is leading and which one is lagging ($Cov(\mu_{jkt}, \mu_{jm,t+s}) = Cov(\mu_{jk,t+s}, \mu_{jmt}) = \sigma_{\mu_k\mu_m,s}$). Fourth, classroom shocks to different student subgroups may be correlated within year ($\sigma_{\theta_k\theta_m}$). Last, the variances of student-type-specific teacher effects ($\sigma^2_{\mu_k} \equiv \sigma_{\mu_k\mu_k,0}$), classroom shocks ($\sigma^2_{\theta_k}$), and individual-level shocks ($\sigma^2_{\varepsilon_k}$) may vary by student type but are constant across years.

[2] A concern about Assumption 1 is that the variance-covariance structure of teacher effects may be different for novice and experienced teachers since teachers' effectiveness rapidly grows early in their careers and becomes stable after three to five years (Rockoff, 2004; Papay and Kraft, 2015). I control for teaching experience when estimating the heterogeneous VA model as a robustness check and find qualitatively similar results.

3.3 Teacher comparative advantage measure

For simplicity and without loss of generality, use the notation $k(i) \in \{0, ..., K-1\}$, and choose student type $k = 0$ as the reference or target group. For the other student types, define D_{ki} as an indicator equal to one if student i belongs to group k and equal to zero otherwise, for $k = 1, ..., K-1$. Equation 2 can be expressed as

$$\nu_{ikt} = \mu_{j0t} + \sum_{\kappa=1}^{K-1} D_{\kappa i} \underbrace{(\mu_{j\kappa t} - \mu_{j0t})}_{CA_{j\kappa t}} + \theta_{ck} + \varepsilon_{ikt}, \tag{3}$$

where the first term, μ_{j0t}, is teacher j's effect on the reference group and the second term, $CA_{jkt} = \mu_{jkt} - \mu_{j0t}$, is the *added* effect on students of type k. While the reference-group-specific VA affects all students, the second term appears as a match effect that only affects type-k students. I refer to the second term as teacher j's revealed CA or teacher CA because it is revealed by the observed disparate impacts on student outcomes.[3] A positive value of teacher CA means that teacher j has a relative advantage or is more effective at teaching type-k students relative to the average teacher, and a negative value means that she has a relative disadvantage for type-k students.[4]

3.4 Estimation of teacher CA

The approach to estimating teacher CA closely follows Chetty et al. (2014a), and I extend it to the more general case of multiple correlated teacher effects. I use prior years of data, excluding year t, to make predictions of student-type-specific teacher VA for year t. Predictions of type-k-specific teacher VA are based not only on information from type-k students but also on information from other student subgroups.[5] I describe the estimation strategy below under the simplifying scenario

[3] I borrowed the term "revealed comparative advantage" from Balassa (1965), who applies this concept to estimate countries' CAs in exporting different goods.

[4] The teacher CA measure can also be interpreted as the degree to which teachers specialize in teaching type-k students. To illustrate this, suppose there are two student types, $k \in \{0, 1\}$, and two teachers who are equally effective with type-0 students but one is more effective with type-1, implying that she has a comparative advantage for type-1 students. Also suppose that there are two similar classrooms except one only has type-0 students, while the other has only type-1 students. To maximize output, one would allocate the teacher with the highest CA to the classroom with the largest proportion of type-1 students and the other teacher to the other classroom. This is the classic example of the Ricardian model of international trade where two countries gain from trade by specializing in the good they have comparative advantage in.

[5] This prediction procedure has some benefits. First, because I only observe one classroom per teacher each year, I cannot separately identify teacher effects (μ_{jkt}) from classroom shocks (θ_{ck}) using year t data alone. Forecasts based on prior years of data overcome this challenge given that classroom shocks are not correlated across time. Second, for the teacher turnover quasi-experiment, I omit test scores from years t and $t-1$ when predicting teacher CA to

that each teacher teaches one classroom per year, teachers have students of every type in their classrooms, class size and classroom composition of students are the same across classrooms and years, and each teacher has the same number of years with available data. Empirically, I account for differences in class size, classroom composition, and number of years with available data.

I first residualize test scores from observable characteristics, $A_{it} = A^*_{it} - X'_{it}\hat{\beta}_k$, where $\hat{\beta}_k$ comes from separate OLS regressions of test scores on observable characteristics using within-teacher variation. To obtain $\hat{\beta}_k$, I split the sample by student type and estimate the following regression for each subsample:

$$A^*_{it} = X'_{it}\beta_k + \alpha_j + \epsilon_{it}, \tag{4}$$

where α_j is teacher fixed effects.

I then construct student-type-specific classroom-level residuals, $\bar{A}_{jkt}$, by taking the average of test score residuals within types for each teacher's classroom in every year:

$$\bar{A}_{jkt} = \frac{1}{n_k} \sum_{i \in \{i: j(c(i,t))=j \ \& \ k(i)=k\}} A_{ikt}, \tag{5}$$

where n_k is the number of students belonging to group k in teacher j's classroom.[6] For example, I compute race-specific classroom averages by taking the average of black students' residualized test scores for each teacher in every year and separately for non-black students' residualized test scores. These classroom-level averages are noisy measures of teacher effects as they include student-type-specific teacher effects and classroom shocks ($\mu_{jkt} + \theta_{ck}$).

I stack the history of teacher j's classroom averages for a given student type into a $(t-1) \times 1$ vector that excludes year t, $\boldsymbol{A}^{-t}_{jk} = \left(\bar{A}_{jk1}, \ldots, \bar{A}_{jk,t-1}\right)'$, and then stack these student-type-specific vectors to form a $K(t-1) \times 1$ grand vector, $\boldsymbol{A}^{-t}_{j} = (\boldsymbol{A}'^{-t}_{j1}, \ldots, \boldsymbol{A}'^{-t}_{jK})'$. Next, I make out-of-sample predictions using prior years' data. Let $\boldsymbol{A}_{jt} = (\bar{A}_{j1t}, \ldots, \bar{A}_{jKt})'$ denote a $K \times 1$ vector that groups teacher j's classroom residuals across student types in year t. The prediction of teacher j's student-

avoid any mechanical correlation between current changes in student quality and changes in estimated teacher quality when teachers move across schools and grades. Third, making predictions using information from all student types help improve precision.

[6]If teachers teach multiple classrooms, one would first compute student-type-specific classroom-level averages and then collapse these classroom residuals at the teacher level using precision weights. The precision weights for classroom c and student-type k would be $h_{ckt} = \frac{1}{\sigma^2_{\theta_k} + \frac{\sigma^2_{\epsilon_k}}{n_k}}$.

type-specific VA for year t, $\hat{\boldsymbol{\mu}}_{jt} = (\hat{\mu}_{j1t}, ..., \hat{\mu}_{jKt})'$, is the best linear predictor of $\boldsymbol{A}_{jt}$ based on prior classroom scores $\boldsymbol{A}_j^{-t}$. It is given by

$$\hat{\boldsymbol{\mu}}_{jt} \equiv \mathbb{E}[\boldsymbol{A}_{jt} | \boldsymbol{A}_j^{-t}] = \boldsymbol{\psi}' \boldsymbol{A}_j^{-t}, \tag{6}$$

where $\boldsymbol{\psi}$ is a $K(t-1) \times K$ matrix of reliability weights. The reliability weights are chosen to minimize the mean-squared error of test scores forecasts:

$$\boldsymbol{\psi} = \arg\min \sum_j \left(\boldsymbol{A}_{jt} - \boldsymbol{\psi}' \boldsymbol{A}_j^{-t} \right)' \left(\boldsymbol{A}_{jt} - \boldsymbol{\psi}' \boldsymbol{A}_j^{-t} \right). \tag{7}$$

Intuitively, under the simplifying scenario, the resulting coefficients are equivalent to those obtained from OLS regressions of $\boldsymbol{A}_{jt}$ on $\boldsymbol{A}_j^{-t}$.[7]

The resulting reliability weight matrix has the form

$$\boldsymbol{\psi} = \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}, \tag{8}$$

where $\boldsymbol{\Gamma}$ is the $K(t-1) \times K(t-1)$ variance-covariance matrix of $\boldsymbol{A}_j^{-t}$ and $\boldsymbol{\gamma}$ is the $K(t-1) \times K$ covariance matrix between $\boldsymbol{A}_j^{-t}$ and $\boldsymbol{A}_{jt}$. In particular, the k-th column of $\boldsymbol{\psi}$, $\boldsymbol{\psi}_k$, which contains the reliability weights to predict type-k-specific teacher VA, $\hat{\mu}_{jkt}$, is

$$\boldsymbol{\psi}_k = \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}_k, \tag{9}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \cdots & \boldsymbol{\Gamma}_{1K} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}_{K1} & \cdots & \boldsymbol{\Gamma}_{KK} \end{pmatrix},$$

[7] The minimization problem has a system of K equations, where the dependent variables are $\bar{A}_{j1t}, \ldots, \bar{A}_{jKt}$ and the explanatory variables are $\bar{A}_{j1,1}, \ldots, \bar{A}_{j1,t-1}, \ldots, \bar{A}_{jK,t-1}$, which are the same variables across equations.

its mn-th block matrix $\mathbf{\Gamma}_{mn}$ is a $(t-1) \times (t-1)$ symmetric matrix

$$\mathbf{\Gamma}_{mn} = \begin{pmatrix} Cov(\bar{A}_{jm1}, \bar{A}_{jn1}) & \cdots & Cov(\bar{A}_{jm1}, \bar{A}_{jn,t-1}) \\ \vdots & \ddots & \vdots \\ Cov(\bar{A}_{jm,t-1}, \bar{A}_{jn1}) & \cdots & Cov(\bar{A}_{jm,t-1}, \bar{A}_{jn,t-1}) \end{pmatrix},$$

and

$$\boldsymbol{\gamma}_k = \begin{pmatrix} \boldsymbol{\phi}_{k,1} \\ \vdots \\ \boldsymbol{\phi}_{k,K} \end{pmatrix},$$

with its m-th element being a $(t-1) \times 1$ vector $\boldsymbol{\phi}_{k,m} = (Cov(\bar{A}_{jkt}, \bar{A}_{jm1}), \ldots, Cov(\bar{A}_{jkt}, \bar{A}_{jm,t-1}))'$.

I use the reliability weights matrix to predict the vector of teacher effects for period t, $\hat{\boldsymbol{\mu}}_{jt}$, following Eq. 6.[8] Finally, I use these student-type-specific teacher VA predictions to construct each teacher's CA, based on Eq. 3. The prediction of teacher j's CA for type-k students (relative to a reference group) is $CA_{jkt} = \hat{\mu}_{jkt} - \hat{\mu}_{j0t}$.

3.5 Identification of population parameters that govern teacher effects

The reliability weights matrix is a function of the variance and covariance parameters. Let $[\mathbf{\Gamma}_{mn}]_{ss'}$ denote the element in the s-th row and s'-th column of the matrix $\mathbf{\Gamma}_{mn}$. Under the stationarity assumptions, the variance and covariance parameters are identified as

$$\begin{aligned}
[\mathbf{\Gamma}_{mm}]_{ss} = Var(\bar{A}_{jms}) &= \sigma^2_{\mu_m} + \sigma^2_{\theta_m} + \frac{\sigma^2_{\varepsilon_m}}{n_m} \\
[\mathbf{\Gamma}_{mm}]_{ss'} = Cov(\bar{A}_{jms}, \bar{A}_{jms'}) &= \sigma_{\mu_m,|s-s'|} \\
[\mathbf{\Gamma}_{mm'}]_{ss} = Cov(\bar{A}_{jms}, \bar{A}_{jm's}) &= \sigma_{\mu_m\mu_{m'}} + \sigma_{\theta_m\theta_{m'}} \\
[\mathbf{\Gamma}_{mm'}]_{ss'} = Cov(\bar{A}_{jms}, \bar{A}_{jm's'}) &= \sigma_{\mu_m\mu_{m'},|s-s'|} \\
Var(A_{ims} - \bar{A}_{jms}) &= \left(1 - \frac{1}{n_m}\right)\sigma^2_{\varepsilon_m}
\end{aligned}$$

[8] The resulting weights are signal-to-noise ratios, and each student-type-specific prediction is shrunk toward its grand mean of zero. Predictions are leave-one-out forecast. They can also be interpreted as empirical Bayes estimators under the distributional assumption that student-type-specific teacher effects follow a multivariate normal distribution, student-type-specific classroom shocks follow a multivariate normal distribution, and individual-level shocks follow a normal distribution.

for all $m \neq m' \in \{1, ..., K\}$ and $s \neq s' \in \{1, ..., t-1\}$, where n_m is the number of students belonging to group m in each teacher's classroom. Note that I cannot separately identify the variance $\sigma^2_{\mu_k}$ from $\sigma^2_{\theta_k}$ and the covariance $\sigma_{\mu_k \mu_m}$ from $\sigma_{\theta_k \theta_m}$.

3.6 Special cases of the DVA model

Previous VA models are special cases of the DVA model. In Appendix Section B, I relate the DVA model to various VA models proposed by Chetty et al. (2014a), Kane and Staiger (2008), Lefgren and Sims (2012), Raudenbush and Bryk (2002), and others. These models set restrictions on certain parameters of the DVA model. For example, Raudenbush and Bryk (2002)'s hierarchical model is the DVA model under the scenario that $\mu_{jkt} = \mu_{jk}$ and $\sigma_{\theta_m \theta_n} = 0$; that is, student-type-specific teacher effects are fixed but correlated with each other, and student-type-specific classroom shocks are uncorrelated with each other.

3.7 Limitations and threats to identification

The DVA model flexibly estimates student-type-specific teacher VA (μ_{jkt}) without imposing any restrictions on their distributions; however, this flexibility comes with a cost for high-dimensional or continuous student types. To semi-parametrically estimate teacher VA across student types, the DVA model splits the sample into subgroups, which may cause small sample problems for high-dimensional student types because there may not be enough students with a particular type in a teacher's classroom. For example, if we are considering student race, gender, and SES types simultaneously, a teacher may only have a handful of poor non-black boys in her classroom to reliably estimate her race-gender-SES-specific VA. Similarly, if we are considering continuous student types, such as achievement level (based on baseline test scores), not all student types would be present in a teacher's classroom.

One way to overcome small sample problems when considering high-dimensional student types is to parameterize teacher effects and impose distributional assumptions on these effects. For example, Ahn et al. (2024) parameterize teacher effects as a linear function of student characteristics and assume these effects are distributed as a mixture of normals (improving on the normality assumption often imposed by traditional VA models). Another way to overcome small sample problems is to incorporate data from other student outcomes, such as socioemotional skills, that may be affected by cognitive skills, to help improve precision in predicting teacher impacts on cognitive skills across subsets of students. A different solution to overcome small sample problems may be to select a higher level of observation, such as schools rather than teachers, to have more students.

Estimates of teacher CA using observational data may be biased due to non-random sorting of students to teachers. If principals differentially sort students from different subgroups to classrooms based on an unobservable trait, this could attenuate or exacerbate the extent of differential effectiveness. For example, if highly-motivated black students are systematically assigned to teacher A instead of teacher B, it might appear that teacher A is more effective or has a comparative advantage for black students, when in reality the high motivation of black students in her class is a contributing factor to their higher achievement. The quasi-experimental strategy aims to address this concern.

4 Quasi-experimental test for student-teacher match quality

4.1 Ideal experiment

Without loss of generality, suppose we are interested in teacher impacts on racial achievement gaps. The ideal experiment to test for the existence of student-teacher match quality would randomly assign both black and non-black students to high- and low-CA teachers, where a teacher's black CA is her *predicted comparative advantage* for black students measured using observational data (before the experiment). Then, one would test whether black students perform differentially better than non-black students after being randomly assigned to a high-CA teacher and perform differentially worse than those who were randomly assigned to a low-CA teacher. In the absence of such an experiment, I employ a quasi-experimental design where black and non-black students are exposed to teachers with varying degrees of CA due to plausibly exogenous exit and entry of teachers.

4.2 Quasi-experimental strategy

Building on Chetty et al. (2014a), I leverage teacher turnover quasi-experiment to test for the existence of student-teacher match effects. Specifically, I test whether the teacher CA measure accurately predicts teachers' disparate impacts on student outcome gaps. Teachers move across schools and across grades within schools, and therefore students are exposed to teachers with varying degrees of CA while progressing through school. The teacher turnover quasi-experiment works as follows. Suppose a school has two racially mixed classrooms in 4th grade and each classroom has a teacher. Also suppose that one teacher is replaced by another who is otherwise similar (same reference-group VA) but has a 2σ higher *predicted* black CA. Because the new teacher is one half of the faculty, this change should cause a 1σ increase on average in 4th-grade black students' *actual* test scores but no change in 4th-grade non-black students' test scores.

The empirical strategy exploits changes in teacher quality from $t-1$ to t; therefore, teacher CA predictions are constructed by excluding data from years t and $t-1$ to avoid any mechanical correlation between changes in teacher quality and actual changes in student quality. More formally, let A_{ksgt} be the average score of students from type k in school s and grade g in year t. Let $\hat{\mu}_{j0t}^{-\{t,t-1\}}$ and $\hat{\mu}_{j1t}^{-\{t,t-1\}}$ be two-year-leave-out forecasts of teachers' non-black- and black-specific VA, respectively. Hence, the two-year-leave-out forecast of teacher CA is $CA_{jt}^{-\{t,t-1\}} = \hat{\mu}_{j1t}^{-\{t,t-1\}} - \hat{\mu}_{j0t}^{-\{t,t-1\}}$ and the forecast for the reference-group VA is $VA_{jt}^{-\{t,t-1\}} = \hat{\mu}_{j0t}^{-\{t,t-1\}}$. Let CA_{sgt} be the average of $CA_{jt}^{-\{t,t-1\}}$ across all teachers in school s and grade g in year t. Similarly, let VA_{sgt} be the school-grade average of $VA_{jt}^{-\{t,t-1\}}$.

I estimate the following empirical model by student type:

$$\Delta A_{ksgt} = \alpha + \lambda_k \Delta CA_{sgt} + \phi_k \Delta VA_{sgt} + \Delta\varepsilon_{ksgt}, \tag{10}$$

where $\Delta A_{ksgt} = A_{ksgt} - A_{ksg,t-1}$ is the cross-cohort change in type-k students' test scores within the school-grade cell, and $\Delta CA_{sgt} = CA_{sgt} - CA_{sg,t-1}$ is the yearly change in average teacher CA quality. Similarly, $\Delta VA_{sgt} = VA_{sgt} - VA_{sg,t-1}$ is the yearly change in average teacher reference-group VA, and $\Delta\varepsilon_{ksgt}$ is the cross-cohort change in unobserved determinants of student achievement. The coefficient of interest is λ_k, which tells us how much changes in teacher CA when a teacher is replaced by another predicts changes in black or non-black students' test scores. Given that black CA is constructed as the added effect on black students, one would expect $\lambda_k = 0$ for non-black students and $\lambda_k = 1$ for black students.

To increase power to test for match effects, I pool the sample of black and non-black students and estimate the following modified regression:

$$\Delta A_{ksgt} = \alpha + \lambda \Delta (D_{ksg} \times \Delta CA_{sgt}) + \phi \Delta VA_{sgt} + \Delta \varepsilon_{ksgt}, \tag{11}$$

where $\Delta A_{ksgt}, \Delta CA_{sgt}, \Delta VA_{sgt}$, and $\Delta \varepsilon_{ksgt}$ are the same as above, D_{ksg} is an indicator equal to 1 if the cohort of students is black and 0 otherwise, and $D_{ksg} \times \Delta CA_{sgt}$ is the matched effect that is equal to ΔCA_{sgt} for cohorts of black students in school s and grade g and equal to zero for cohorts of non-black students.[9] If $\lambda = 1$, then teacher CA accurately predicts teachers' disparate impacts on student outcome gaps. Here, $B = 1 - \lambda$ is the degree of forecast bias (Chetty et al., 2014a).

The following identifying assumption is necessary to interpret results as causal:

Assumption 2 (Conditional independence in teacher turnover quasi-experiment) *Changes in teacher CA across cohorts within a school grade are orthogonal to student-type-specific changes in unobserved determinants of student achievement, conditional on changes in general teacher quality.*

$$\Delta CA_{sgt} \perp \Delta \varepsilon_{ksgt} | \Delta VA_{sgt}. \tag{12}$$

The identifying Assumption 2 tells us that changes in teacher CA quality is uncorrelated with unobserved student-type-specific determinants of student achievement, after controlling for changes in overall teacher quality. This assumption would be violated by student-type-specific non-random sorting of students and teachers over time, for example, if black students follow high-CA teachers across schools, or if student quality is differentially changing by race within a school and teachers sort to these schools based on their CA. Although possible, this race-specific sorting seems unlikely as the cost for parents to transfer their children to another school may prevent them from following teachers. Moreover, the quasi-experiment exploits high-frequency yearly changes in teaching staff, and thus student-type-specific sorting would need to occur at a high frequency to invalidate

[9] The more general form of Eq. 11 to accommodate multidimensional student types is:

$$\Delta A_{ksgt} = \alpha + \sum_{\kappa=1}^{K-1} \lambda_\kappa \Delta (D_{\kappa sg} \times \Delta CA_{\kappa sgt}) + \phi \Delta VA_{sgt} + \Delta \varepsilon_{ksgt},$$

where CA_{ksgt} is constructed based on $CA_{jkt}^{-\{t,t-1\}} = \hat{\mu}_{jkt}^{-\{t,t-1\}} - \hat{\mu}_{j0t}^{-\{t,t-1\}}$ for $k = \{1, ..., K-1\}$, which is the two-year-leave-out comparative advantage forecast for type-k students, D_{ksg} is an indicator variable equal to 1 if the cohort of students belongs to type k and 0 otherwise, VA_{sgt} is the reference-group VA based on $VA_{jt}^{-\{t,t-1\}} = \hat{\mu}_{j0t}^{-\{t,t-1\}}$, and A_{ksgt} is the average score of students from type k.

this strategy (Chetty et al., 2014a). Note that Assumption 2 implies that match effects are also conditionally orthogonal: $(D_{ksg} \times \Delta CA_{sgt}) \perp \Delta\varepsilon_{ksgt} | \Delta VA_{sgt}$.

5 Data

5.1 Chicago Public Schools data

I use de-identified administrative data from Chicago Public Schools to test whether teacher quality has a student-teacher match component. These data span from the 2008–09 to 2016–17 school years and contain information on student test scores, student demographics, attendance, suspensions, and course transcripts. These data also contain information on teacher characteristics, classroom observation ratings, and student surveys.

Test scores. Two different exams were administered during the period of analysis, namely the Illinois Standards Achievement Test (ISAT) and the Northwest Evaluation Association Measures (NWEA). ISAT was administered until 2013–14 to students in grades 3–8 in public schools and measured achievement in math, English language arts (henceforth "English"), and science (for grades 4 and 7). NWEA has been administered since 2012–13 to students in grades 2–8 to assess math and English skills. In the first year of administering NWEA, a baseline assessment was given in the fall of 2012, which I treat as 2011–12 test scores. ISAT and NWEA overlap in three school years (2011–12, 2012–13. and 2013–14), and I select NWEA for these years because it has a larger number of test takers.

Math and English test scores are the main outcomes. I normalize test scores to have mean 0 and standard deviation 1 by subject-grade-school year.

Student characteristics. Data include race/ethnicity, gender, age, reduced and free lunch status, special education status, attendance, and suspensions.[10] Ethnicity is treated as race in the data, and I cannot distinguish between black Hispanic students from their non-black Hispanic peers if they selected Hispanic.[11]

Teacher characteristics. I use personnel data to relate teacher characteristics with teacher CA

[10] Each student has a fixed value for their gender, race/ethnicity, and date of birth. If any of these characteristics change across years for any given student, I assign the mode. Age is computed as of September 1 of each year.

[11] If black Hispanics chose the Hispanic option rather than African American/black, this could underestimate the importance of race (i.e., the variance of teacher CA for black students) because their outcomes tend to be more similar to those of African Americans than to non-black Hispanics (Holder and Aja, 2021).

in supplemental analyses. These data include race/ethnicity, gender, age, teaching experience, educational level, and tenure status.[12]

Classroom observations. I use classroom observation data to study teacher behavior and its relationship with teacher CA. In the 2012–13 school year, Chicago Public Schools implemented a new district-wide teacher evaluation system, called Recognizing Educators Advancing Chicago's Students. Under this system, evaluators, mainly school administrators, observe teachers in the classroom multiple times a year and rate them on various teaching practices rubrics that are based on the Danielson Framework for Teaching. These practices are clustered into four domains: planning and preparation (domain 1), classroom environment (domain 2), instruction (domain 3), and professional responsibilities (domain 4). Each domain in turn is formed by components, and each component is rated on a scale from 1 to 4, where 1 means unsatisfactory and 4 distinguished.[13] I use the aggregated scores at the domain level in the analysis.

Student surveys. I use student surveys to relate students' ratings of their teachers with their teachers' CA. Data come from Chicago Public Schools 5Essentials student survey, which has been administered annually to students from grades 6 to 12 and has been used as a metric for school quality. The survey includes questions about students' experiences with peers and teachers, attitudes, and activities in school. Completion rates are high, with about 80 percent, and the data I use span from 2011–12 to 2016–17. I extract 36 survey items that were consistently asked across years and were about the students' math and English teachers and courses. I follow the survey administrator's theoretically-informed grouping of questions and construct seven indexes (peer support, classroom rigor, academic press, course clarity, academic engagement, academic professionalism, and classroom disruptions) by taking the average of their respective items' scores.[14] Appendix Figure A.2

[12] I construct teaching experience based on the last hiring date or as the cumulative number of years that the teacher appears in the data if the hiring date is missing.

[13] The domains and their components are the following. Domain (1) planning and preparation: component 1a) demonstrating knowledge of content and pedagogy, 1b) demonstrating knowledge of students, 1c) selecting learning objectives, 1d) designing coherent instruction, and 1e) designing student assessment. Domain (2) classroom environment: component 2a) creating an environment of respect and rapport, 2b) establishing a culture for learning, 2c) managing classroom procedures, and 2d) managing student behavior. Domain (3) instruction: component 3a) communicating with students, 3b) using questioning and discussion techniques, 3c) engaging students in learning, 3d) using assessment in instruction, and 3e) demonstrating flexibility and responsiveness. Domain (4) professional responsibilities: component 4a) reflecting on teaching and learning, 4b) maintaining accurate records, 4c) communicating with families, 4d) growing and developing professionally, and 4e) demonstrating professionalism. Appendix Figure A.1 shows an example for the rubric used to rate component 2a) creating an environment of respect and rapport.

[14] The short description of each index is as follows. Classroom rigor: whether or not teachers encourage all students to make connections and seek multiple perspectives through their coursework. Academic press: asks students' views

lists the survey items that form each index.

Link of students to teachers. I use transcript records to link students to their teachers. These data provide detailed course-taking information for each student, including courses the student was enrolled in, course grades, and identifiers for the teachers who provided the final grade. I identify each student's math and English courses and their respective teachers.[15] A limitation of the data is that I do not observe actual classroom assignments. For example, if a teacher teachers two sections of the same course, I cannot determine which students belong to each section. Therefore, I construct new classroom rosters by pooling all students who are in the same school and linked to the same math or English teacher. I refer to these newly constructed rosters as classrooms for convenience. As a result, each teacher has one classroom per year by construction.

5.2 Sample selection

I organize the dataset to have one row per student per subject (math and English) per school year and perform similar sample restrictions as prior work. This implies that each subject-teacher is a different treatment. I first restrict the sample to students in grades 3–8, where prior test scores are available,[16] and I am left with 2.4 million math and English test scores. Next, if a teacher teaches in multiple schools in a year, I keep the links of one school and set the other links to missing because the quasi-experiment requires one teacher per school (1.4 percent of observations).

I then remove classrooms that have more than 25 percent of students with special needs because they may have multiple teachers (8.4 percent of observations). Out of the remaining classrooms, I drop those with less than seven students to have a sufficient number of students to estimate teacher effects, and I also drop teachers linked to more than 200 students because they could be mislinked or be a dedicated staff that enters the grades and not the students' actual teacher (1.5 percent of

of their teachers' efforts to push students to higher levels of academic performance and teachers' expectations of student effort and participation toward academic achievement. Course clarity: captures students' views about what they need to succeed in the target class, their learning from feedback, and how helpful the homework and class work are. Academic engagement: examines student interest and engagement in learning. Academic personalisms: gauges whether students perceive that their classroom teachers give them individual attention and show personal concern for them. Classroom disruptions: reports on the degree to which students disrupt their learning.

[15]One shortcoming of this linkage is that the teacher who assigned the final grade may not be the student's regular teacher. Although I cannot verify it, this practice may not be very common across all schools. Furthermore, if a school has a dedicated person who submits grades, she would be dropped from the analytic sample (as I describe below) if the number of linked students crosses a threshold.

[16]During the school years 2008–09 to 2011–12 when ISAT was in place, third graders are excluded because they do not have prior test scores.

observations). The remaining data have 2.2 million student-subject-year observations, and this is the core sample used in the quasi-experimental strategy. To estimate teacher CA, I further restrict the sample to students with information on prior and current test scores, demographics, and teacher assignments, leaving me with about 1.8 million student-subject-year observations.

5.3 Descriptive statistics

Chicago Public School students tend to come from low-income families and are overwhelmingly racialized minorities. Table 1, Panel A reports summary statistics of the sample used to estimate teacher effects. I observe 331,369 unique students who appear in the data for 5.4 school years on average. The mean test score is 0.11, higher than 0 because test scores were normalized using the full population, which included special education students who tend to have lower scores. Half of the students are female, and the average age is 10.7. The percentage of black non-Hispanic (henceforth "black") students is 37 percent, and the percentage of Hispanic students is 48 percent. A high percentage of students are eligible for free or reduced-price lunch (86 percent). Only 1 percent of students are repeating the grade. Classrooms in Chicago Public Schools are gender balanced but highly racially segregated. About 20 percent of classrooms has at least seven black students and seven non-black students, while virtually every classroom (98 percent) has both girls and boys.[17] The average number of students in each teacher-subject-year cell is 39.

[Table 1 about here]

Descriptive statistics of teachers are reported in Appendix Table A.1. These teacher characteristics are not used in estimating teacher effects but in supplemental analysis about sources of heterogeneity in teacher CA. I observe 9,740 unique teachers and 15,894 unique teacher-subject cells that appear in the data for three years on average. The majority of teachers are female (85 percent), 46 percent are white, 29 percent are black, and 20 percent are Hispanic. About 20 percent of teachers are ever observed teaching at least seven black and seven non-black students in their teaching career. Nonetheless, 50 percent of teachers are ever observed teaching to at least seven black students and 70 percent have taught at least seven non-black students.[18]

[17] As I mention in Section 6.2, a classroom must have at least seven students from one type to be included in the DVA model estimation.

[18] A teacher is said to have ever taught black students if any of her classrooms in the present or the past have had

6 Empirical application: Teacher impacts on racial achievement gaps

In this section, I apply the DVA model to study teacher impacts on racial achievement gaps in Chicago Public Schools. The next section formally tests for student-teacher match effects and investigates whether the documented racially disparate impacts are causal. Black students tend to score lower than other students; therefore, I focus on teacher impacts on black versus non-black students. The non-black racial group includes Hispanic students as well as the small percentage of white students and students from other races. For simplicity, I use two student racial types but the DVA model can be extended to have multiple types.

6.1 Racial achievement gaps

To fix ideas about the extent to which teachers impact racial achievement gaps, Table 2 documents racial test score gaps for Chicago Public Schools. Column 1 regresses student test scores (math and English) on race/ethnicity indicators and shows an unconditional black-white achievement gap of 0.88σ and a smaller Hispanic-white achievement gap of 0.71σ. There are no statistically significant gaps between white students and students from other races. Column 2 further controls for prior year's test scores and shows a conditional black-white achievement gap of 0.18σ and a Hispanic-white achievement gap of 0.14σ.

[Table 2 about here]

Columns 3 and 4 report the black-non-black gap by regressing student test scores on the black race indicator. The unconditional black-non-black achievement gap is 0.35σ (Column 3) and the conditional gap is 0.08σ, both of which are lower than the respective black-white achievement gap due to the inclusion of Hispanic students in the reference group. Throughout the paper, I refer to the unconditional black-non-black achievement gap as racial achievement gap, as it matches with the two analyzed racial groups and is more conservative than the conditional gap. Looking at the racial achievement gaps by subject and school level, the largest unconditional test score gap is in elementary school math, where the black-non-black gap is 0.41σ (see Appendix Table A.2).

at least seven black students. This is similarly defined for ever teaching non-black students and other demographics.

6.2 Estimation and output of the DVA model

I estimate race-specific teacher VA separately by subject (math and English) and school level (elementary and middle school). Following the steps described in Section 3.4, I first split the sample by race (black and non-black) and residualize test scores with respect to covariates. I restrict the sample to classrooms with seven or more students from type k to reduce the influence of small sample problems.[19] The set of covariates, X_{it}, is similar to that used in prior work and includes student characteristics, classroom- and school-level averages of these characteristics, and grade and year fixed effects.[20] Additionally, I include black- and non-black-specific cubics in class mean prior test scores (e.g., class mean average of black students' test scores) and the number of black and non-black students in the classroom to control for race-specific peer effects. For each subsample, I estimate the parameter β_k using within-teacher variation.[21] With the estimates $\hat{\beta}_k$, I construct (precision-weighted) classroom-level averages of individual-level residuals separately for each student type, $\bar{A}_{jkt}$.

Next, I use the history of race-specific classroom-level residuals to estimate the variance-covariance matrix of teacher effects as well as the variance of classroom- and individual-level shocks (discussed in Section 3.5). Then, I construct the reliability weights matrix and make forecasts of race-specific teacher VA based on prior years' data and the reliability weights matrix. Last, I construct teacher CA predictions as the difference between black- and non-black-specific teacher VA forecasts. Table 3 reports the output of the DVA model.

Auto- and cross-covariance matrix. Is information on non-black students' performance informative about teacher impacts on black students? Panels A and B of Table 3 present estimates of the variance-covariance matrix and correlations (in brackets) for elementary and middle school levels, respectively. Columns 1–3 show estimates for math and Columns 4–6 for English. Column 1 shows the autocovariance vector of black-specific teacher VA, Column 2 shows the autocovariance

[19] If a classroom has 10 students of the same type, they are included in regressions to compute residualized test scores. However, if the classroom has 5 students of each type, they are dropped from the regressions.

[20] Covariates include the following: (i) cubic polynomials in prior test scores in math and English, interacted with grade; (ii) student's gender, race/ethnicity, age, free or reduced-price lunch status, special education status, grade repetition indicator, indicator if baseline test score is below the respective subject and grade mean, logarithm of baseline number of absences, in-school suspension and out-of-school suspension (log of variable plus one); (iii) class and school-year means of student characteristics; (iv) cubics in class and school-grade mean prior test scores, each interacted with grade; (v) teacher-subject-year cell size (class size); and (vii) grade and year fixed effects.

[21] Including or excluding teacher fixed effects does not meaningfully change the results. The correlation of residuals with and without teacher fixed effects is 0.99.

of non-black-specific teacher VA, and Column 3 shows the cross-year covariance between black- and non-black-specific teacher VA, which is symmetric as to which variable is leading or lagging. Columns 4–6 follow the same structure for English.

[Table 3 about here]

Figure 1 shows the graphical representation of the auto-correlations (depicted as black circle and red triangle lines) and cross-correlations (blue diamond line) by subject and school level. Three main points emerge from Figure 1. First, the autocorrelations and cross-correlations decay over time, suggesting that prior years' VA are more informative in predicting the current year's VA than later ones. Second, the autocorrelation and cross-correlation vectors for math are shifted upward relative to English, which indicates a higher persistence of teacher effects for math. Last and importantly, the cross-correlation vectors are close to the autocorrelation vectors, meaning that a teacher's effect on one student subgroup is informative in predicting her effects on the other student subgroup. I use these auto- and cross-correlation vectors to make forecasts of black- and non-black-specific teacher VA based on prior years of data.[22]

[Figure 1 about here]

Variances. Are teachers more important for the achievement gains of black or non-black students? If the variance in the teacher effects is much larger for black students, then having a teacher in the top (bottom) quartile of effectiveness would be more beneficial (detrimental) for a black student than for a non-black student. Panels C and D of Table 3 report estimates of the within-year population variances and covariances of teacher effects for elementary and middle schools, respectively. Columns 1 and 2 show estimates for math and Columns 3 and 4 for English.

In Column 1 of Panel C of Table 3, the first row is the standard deviation of residualized test scores, the second row is the standard deviation of individual-level shocks for black students, and the third row shows the sum of variances $\sigma^2_{\mu_{black}} + \sigma^2_{\theta_{black}}$. Although it cannot be separately identified because I observe only one classroom per teacher per year, the black-specific variance $\sigma^2_{\mu_{black}}$ can be

[22]Estimates of autocorrelations and cross-correlations are noisy, especially for longer time horizons due to smaller sample sizes. (Appendix Table A.3 presents the number of observations used to calculate the variance-covariance matrix of race-specific teacher VA by subject and school level.) I set the correlations with year gaps greater than four to the corresponding correlation at year gap four when making out-of-sample forecasts; i.e., $\sigma_{\mu_k \mu_m s} = \sigma_{\mu_k \mu_m 4}$ for $s > 4$.

estimated (row 4) by regressing the history of black-specific autocovariance estimates $\{\hat{\sigma}_{\mu_{black},t}\}_{t=1}^{t=7}$ on a quadratic polynomial of time trends ($t = \{1, ..., 7\}$) and then projecting it onto $t = 0$. Column 2 of Panel C presents the same estimates for non-black student type. Results indicate that population variability estimates of black- and non-black specific teacher VA in Chicago are similar and lie within the variability estimates from New York and Los Angeles. For example, the variability of black- and non-black-specific teacher VAs for elementary math both are 0.23σ, while the reported estimate for New York is 0.16σ and Los Angeles is 0.29σ (Chetty et al., 2014a; Bacher-Hicks et al., 2014).[23]

Within-year covariance. To what extent are teacher impacts on black and non-black students correlated? An innovation of the DVA model is that it allows me to estimate the true within-year correlation between teacher effects for different student types. The naive pairwise correlation between student-type-specific teacher VAs (i.e., $Corr(\hat{\mu}_{j0t}, \hat{\mu}_{j1t})$) may not yield the true correlation because each of these teacher VAs is measured with error. The true correlations could be higher or lower due to attenuation bias from random estimation errors and the measurement errors being correlated across student types in the same year (i.e., $\sigma_{\theta_0\theta_1} \neq 0$) (Jackson et al., 2024).

Row 5 in Table 3 Panels C and D shows the sum of the within-year covariances $\sigma_{\mu_{black}\mu_{nonblack}} + \sigma_{\theta_{black}\theta_{nonblack}}$ for elementary and middle schools, respectively. Although it cannot be separately identified because I observe only one classroom per teacher each year, I estimate the within-year covariance $\sigma_{\mu_{black}\mu_{nonblack}}$ by regressing the history of cross-year covariance estimates ($\{\sigma_{\mu_{black}\mu_{nonblack},t}\}_{t=1}^{t=7}$) on a quadratic polynomial of time trends ($t = \{1, ..., 7\}$) and then projecting it onto $t = 0$. The true correlation between black- and non-black-specific teacher VA is 0.89 ($= \hat{\sigma}_{\mu_{black}\mu_{nonblack}}/(\hat{\sigma}_{\mu_{black}} \times \hat{\sigma}_{\mu_{nonblack}}) = 0.048/[0.231 \times 0.233]$) and 0.99 for elementary math and English, respectively, and 0.76 and 0.33 for middle school. These correlations are in most cases lower than the naive correlation estimates, which vary between 0.53 and 0.97 (see Appendix Table A.4 row 6), suggesting that naive correlations understate heterogeneity in teacher impacts across student types.

[23] Chetty et al. (2014a) report estimates of homogeneous teacher VA for New York City and find that the standard deviation in teacher impacts is 0.16σ and 0.12σ for elementary school math and English, respectively, and 0.08σ and 0.13σ for middle school. Similarly, Bacher-Hicks et al. (2014) report estimates of homogeneous teacher VA for Los Angeles and find a standard deviation of 0.29σ and 0.19σ for elementary math and English, respectively, and 0.21σ and 0.10σ for middle school. In Chicago, the true variability of black-specific teacher VA is 0.23σ and 0.17σ for elementary math and English, respectively, and the corresponding figures for non-black-specific teacher VA are 0.23σ and 0.13σ. For middle school, the true variability of black-specific teacher VA is 0.14σ and 0.11σ for math and English, respectively, and the corresponding numbers for non-black-specific teacher VA are 0.14σ and 0.09σ (row 4 of Panels C and D of Table 3). A consistent finding across these three cities and other studies is that the variance of teacher effectiveness is larger for math than English and larger for elementary school than middle school.

6.3 Empirical distribution of teacher CA

Empirical density. Do teachers differentially impact black and non-black students? Setting non-black students as the reference group, I construct a teacher's black CA as the difference between her forecasted black- and non-black-specific teacher VA. Figure 2, Panels A and B plot the empirical distribution of teachers' black CA for elementary and middle schools, respectively.[24] The solid black line in each figure refers to math teachers' black CA, and the dotted red line refers to English teachers. These densities are constructed using the sample restricted to teachers who are ever observed teaching at least seven students from each race and weight each teacher-subject-year observation by the total number of students in the cell. Each graph reports the empirical standard deviation of teacher CA estimates at the top right corner, which varies between 0.04–0.07 standard deviation. These empirical variances are statistically different from zero based on permutation tests.[25]

[Figure 2 about here]

Variance. To what extent do teachers impact racial achievement gaps? The above empirical standard deviations of teacher CA understates the true variation because they are obtained from shrunken teacher VA estimates. The true variance of teacher CA reflects the actual spread of disparate impacts on student outcomes; on the other hand, the shrunken variance is smaller than the true variance due to the statistical adjustment of shrinkage that accounts for uncertainty. To address this, I estimate the true variability of teacher CA as $Var(CA_{jt}) = \hat{\sigma}^2_{\mu_{black}} + \hat{\sigma}^2_{\mu_{nonblack}} - 2\hat{\sigma}_{\mu_{black}\mu_{nonblack}}$, where the within-year variance and covariance estimates are described in the previous section. The true standard deviation of black CA is 0.11σ and 0.04σ for elementary school math and English, respectively, and 0.10σ and 0.12σ for middle school (see Appendix Table A.4 row

[24] Appendix Figure A.3 shows the empirical marginal distribution of black- and non-black-specific teacher VA, by subject and school level, which are the basis for constructing teacher black CA.

[25] I test whether the variability of teacher CA is 0, or equivalently that teacher impacts are homogeneous across student types, separately by subject and school level. Under the null hypothesis of homogeneous teacher impacts, the empirical distribution of the variance of teacher black CA is obtained by first randomly assigning students to black and non-black types (keeping the proportions of black and non-black students intact within classrooms). With the permuted sample, I re-estimate the DVA model to obtain the variance-covariance matrix of race-specific teacher VA and make forecasts of black- and non-black-specific teacher VA. Then I use these new forecasts to construct teacher black CA and obtain its empirical standard deviation. I repeat this exercise 100 times and calculate the p-value as the proportion of simulated standard deviations that are greater than or equal to the observed standard deviation. I reject the null hypothesis at conventional levels for each subject and school level (p-value <0.01).

4). These numbers are larger than the naive standard deviation of shrunken CA estimates reported above (see Appendix Table A.4 row 7).

Replacing an average elementary math teacher by another who is otherwise similar except that she is at the 75th percentile of the black CA quality distribution is predicted to differentially increase black students' test scores by 0.10σ. This effect is equal to 41 percent of the black-non-black achievement gap in elementary math. Across subjects and school levels, this thought experiment would close the black-non-black achievement gap by 15–41 percent (or, by extrapolation, close the black-white achievement gap by 5–15 percent).

Stability. Teacher CA is persistent over time. Its stability, measured as the year-to-year correlation, is 0.71 for elementary math teachers and oscillates between 0.61 and 0.89 across subjects and school levels (see Appendix Table A.4 row 8). The stability of teacher CA is slightly lower than the stability of homogeneous teacher VA in my setting, which ranges between 0.77 and 0.82, but is higher than those reported in other settings, between 0.18 and 0.64 (Koedel et al., 2015), perhaps because the DVA model estimates persistent teacher effects.

7 Quasi-experimental validation of teacher CA

7.1 DVA model fitting

Teacher CA forecasts are best linear predictors based on student test scores; therefore, they should have a one-to-one relationship with student test scores depending on the students' type (Chetty et al., 2014a). To show that the DVA model produces valid forecasts as one should expect, I run the following regression by student type:

$$A_{it} = \alpha + \lambda_{k(i)} CA_{jt} + \phi_{k(i)} VA_{jt} + \varepsilon_{it}, \tag{13}$$

where A_{it} is residualized student test scores of student i, whose type is $k \in \{black, nonblack\}$, in year t; CA_{jt} is her teacher j's black CA prediction for year t based on prior years of data; VA_{jt} is teacher j's reference-group-specific VA predictions; and ε_{it} is the error term. The coefficient of interest is λ_k, which tells us the association between teacher CA forecasts and student test scores for a given a student type k. Regressions include subject- and school-level indicators and cluster standard errors by school-cohort. There is one observation per student-subject-year.

Table 4 confirms that teacher black CA predictions explain black students' test scores but not non-black students' scores, as one would expect. The sample is split by race with the sample of black students in Columns 1–3 and non-black students in Columns 4–6. In Column 1, I regress residualized test scores on black CA, Column 2 on reference-group VA, and Column 3 on both teacher CA and reference-group VA. Columns 4–6 follow the same structure. Column 1 shows that a 1σ increase in black CA predicts a 0.84σ increase in residualized test scores for black students, and Column 2 shows that a 1σ increase in teacher reference-group VA predicts a 0.91σ increase in student test scores. In Column 3, the coefficient for black CA and reference-group VA are 1.09 and 0.97, respectively, and I cannot reject at conventional levels that each coefficient is equal to one in the sample of black students. Columns 4–6 show that teacher black CA has a negative relationship with non-black students' test scores when it is the only explanatory variable, but it becomes indistinguishable from zero when both teacher CA and reference-group VA are simultaneously included in the regression. Moreover, I cannot reject that teacher reference-group VA has a one-to-one relationship with non-black students' test scores.

[Table 4 about here]

These results indicate that teacher CA measure is a match effect that only predicts black students' test scores. Given the non-random sorting of students to teachers, a coefficient of $\lambda_k = 1$ does not necessarily imply causality. In the next section, I employ a quasi-experimental strategy to test whether teacher CA reflects differential teacher effectiveness or sorting.

7.2 Main results of the quasi-experimental test

I start by reproducing the main results of Chetty et al. (2014a), who test for forecast bias of homogeneous teacher VA. I estimate a single teacher effect per teacher based on past data, excluding years t and $t-1$. I estimate a model analogue to Eq. 10, where I regress changes in average test scores at the school-grade-subject level on changes in estimated homogeneous teacher VA at the school-grade-subject level. The results of the replication exercise are in Appendix Table A.6, where Columns 1–6 reproduce Table 4 of Chetty et al. (2014a) and Column 7 their Table 5, Column 1. Their main results and robustness checks hold in my data in that I cannot reject that a homogeneous teacher VA is forecast unbiased.[26] The implied forecast bias in the main specification is 1.6 percent (s.e. = 0.070), while that of Chetty et al. (2014a) is 2.6 percent (s.e. = 0.033).

Next, I test for forecast bias of teacher CA by estimating Eq. 10 and present the results in Figure 3 and Table 5. I regress changes in average test scores at the student type-school-grade-subject level on changes in estimated teacher CA and VA at the school-grade-subject level. Each panel in Figure 3 plots changes in average test scores across cohorts versus changes in teacher quality, where Panels A and B relate black and non-black students' test scores with teacher reference-group VA, and Panels C and D with teacher black CA. Each figure is a binned scatterplot where the sample is grouped into 20 equally sized groups based on the specified x-variable and the average of the y-variable is plotted for each group. While teacher reference-group VA predicts changes in both black and non-black students' test scores (Panels A and B in Figure 3), teacher CA predicts changes in black students' test scores only (Panels C and D). The underlying regression results are in Table 5.

[Figure 3 about here]

Table 5 presents the main results separately for the sample of black students (Columns 1–3) and non-black students (Columns 4–6). Regressions include year fixed effects, weight observations by the number of students in the student-type-school-grade cells, and cluster standard errors by school-cohort. Column 1 controls only for teacher black CA and indicates that a 1σ change in this measure predicts a 0.62σ change in black students' test scores. Column 2 controls only for reference-group teacher VA and shows that a 1σ increase in this measure predicts a 0.47σ increase in student test scores. Column 3 includes both teacher black CA and VA and shows that a 1σ increase in each of these measures predicts a 1.06 and 0.97σ increase in black students' test scores, respectively. I cannot reject that each coefficient is equal to one. For the sample of non-black students, Column 6 shows that a 1σ increase in teacher black CA predicts a -0.07σ change in test scores, while a same magnitude increase in teacher VA predicts a 0.89σ increase in test scores. I cannot reject that the teacher CA coefficient is equal to zero and the teacher VA coefficient is equal to one.

[Table 5 about here]

[26]One exception is the robustness specification that controls for leads and lags of changes in mean teacher VA and the cubic polynomial of change in lagged mean student test scores (Column 3 of Appendix Table A.6. Excluding lagged mean student test scores would change the teacher VA coefficient close to one, and I cannot reject that it is equal to one (results not shown here).

Next, I pool the sample of black and non-black students to increase precision and construct the interaction term $D_{ksg} \times \Delta CA_{sgt}$ to test for match effects. Columns 7–9 of Table 5 present the results of Eq. 11, where the explanatory variables are matched teacher black CA (the interaction term) and teacher reference-group VA. As expected from previous results, in the specification that controls for changes in both teacher quality measures (Column 9), I cannot reject that matched black CA is forecast unbiased. The implied forecast bias is -3.9 percent (s.e. = 0.107). I also cannot reject that teacher reference-group VA is forecast unbiased (implied forecast bias = 6.6 percent, s.e. = 0.070), supporting prior work that shows homogeneous teacher VA is forecast unbiased, given that these two different measures have a one-to-one relationship (see Appendix Table A.5 Column 2).

I then investigate whether the main results hold across various subsamples in Appendix Table A.7 Panel A. Column 1 reproduces the main results for teacher black CA, and Columns 2–5 split the sample by subject and school level. The coefficient of matched teacher black CA is statistically equal to one across subjects and school levels.

The identification of match effects comes from the subset of classrooms with both black and non-black students, which account for about 20 percent of all classrooms in Chicago Public Schools. This may raise concerns about the external validity of the results for the remaining classrooms, which may be systematically different from the racially integrated classrooms. To indirectly address this, I perform the same quasi-experimental strategy using student gender types, instead of race types, given that virtually all classrooms have both girls and boys. Appendix Figure A.4 shows the racial and gender composition across classrooms in Panels A and B, respectively. The proportion of black students is bimodal with peaks at 0 and 1, and the proportion of female students is unimodal centered at 0.5, which means that a large proportion of teachers have classrooms with black students only or non-black students only, while every teacher has classrooms with both girls and boys.

I employ the DVA model to estimate gender-specific teacher VA and construct a teacher's CA for female students as the difference between her female- and male-specific teacher VA. I find that teacher female CA is forecast unbiased with an implied bias of 14 percent (s.e. = 0.204) (see Appendix Table A.7 Panel B). In sum, the quasi-experimental results indicate that teacher CA is a good predictor of teachers' disparate impacts across student types, at least when these types are defined by student race and gender.

7.3 Robustness checks

I evaluate the robustness of the result that teacher CA exhibits little or no forecast bias in Table 6. Column 1 reproduces the main result (from Table 5 Column 9), which only uses prior data to make forecasts. Given that years t and $t-1$ are excluded in the two-year-leave-out forecasts, these predictions are only available to teachers with at least three years of data. To increase precision, prior work, including Chetty et al. (2014a), makes jackknife predictions based on prior and future years of data excluding year t (i.e., they additionally use information from $t+1$, $t+2$, etc.). Column 2 follows these studies and shows an insignificant forecast bias with a smaller standard error (implied bias = 6.5 percent, s.e. = 0.072).[27]

One concern that would invalidate the identification Assumption 2 is that improvements in teacher CA may be correlated with other improvements in a school that also differentially increase test scores, for example, if a school makes its climate more inclusive and safe for black students or implements a more inclusive curriculum. Column 3 of Table 6 addresses this concern by replacing year fixed effects with school-by-race-by-year fixed effects to control for any changes that occur within a school that differentially affect test scores across student types. I cannot reject that matched teacher CA is forecast unbiased. Column 4 further controls for trends in quality by controlling for lag and lead changes in matched teacher CA and in teacher VA, as well as cubics in the change across cohorts in prior-year student-type-specific mean own-subject and other-subject test scores. In this specification, teacher CA becomes forecast biased, but after excluding changes in prior-year student-type-specific mean scores, it becomes statistically equal to one (Column 5).[28]

[Table 6 about here]

Placebo tests yield effects of teacher CA that are statistically different from one. Columns 6 and 7 of Table 6 use changes in test scores of the other subject (i.e., English scores for math teachers and math scores for English teachers) for elementary and middle school levels, respectively. The coefficient of teacher CA in these specifications is statistically different from one for elementary and

[27] Under the stationarity Assumption 1, the variance-covariance structure of teacher effects remains the same whether I use past data to make predictions or use past and future data.

[28] The same pattern is observed in Appendix Table A.6, where I replicate prior studies (Chetty et al., 2014a) for univariate teacher VA. In this replication, the homogeneous teacher VA coefficient forecast biased after controlling for changes in prior-year mean test scores across cohorts, in addition to school-by-year fixed effects and lagged and lead changes in teacher quality (Column 3). However, it becomes forecast unbiased after excluding the lagged scores controls (Column 4).

middle schools and is statistically equal to zero for middle schools.[29] Column 8 uses changes in lagged test scores as another placebo test. I find a coefficient of teacher CA that is different from one but also different from zero. The positive association between teacher CA and lagged test scores may be due to teacher effects being estimated from prior years' test scores (Chetty et al., 2017).

The quasi-experimental strategy exploits changes in teacher quality due to teachers moving across grades within, entering, and leaving a school. Column 9 of Table 6 employs an IV strategy where changes in teacher CA and VA are instrumented by changes in the average quality of teachers who leave the school. Changes due to departures are likely to be uncorrelated with high-frequency changes in student quality across cohorts. I find a forecast bias of matched teacher black CA that is indistinguishable from zero. In sum, robustness checks suggest that the results are robust to various specifications and support the finding that teacher CA accurately predicts the effects of teachers on student achievement gaps.

7.4 Multidimensional student types: gender, socioeconomic status, and baseline achievement level

Prior work has documented important disparities in teacher value-added across multiple student types, including gender, socioeconomic status, and baseline achievement level.[30] I apply the DVA model and quasi-experimental strategy to test for student-teacher match quality across these dimensions of student heterogeneity. I use gender to create two student types: girls and boys. In parallel, I use free/reduced-price lunch status as a measure of students' socioeconomic status to create two student types: poor and non-poor. Additionally, I use prior-year test scores to create two student types: low and high achievers, where the lower-achieving group scored below the mean for their respective subject and grade level and the higher-achieving group scored above.[31] One could potentially create more than two student groups (e.g., low, medium, and high achievers) or

[29] This result coincides with Chetty et al. (2014a), who find the coefficient of homogeneous teacher VA to be different from one and zero for elementary schools and different from one but equal to zero for middle schools.

[30] Another student characteristic available in the data is special education status, but it is not discussed here because the sample selection drops classrooms with a large proportion of students with special needs.

[31] Appendix Figure A.4 shows the distribution of the proportion of black students (Panel A), female students (Panel B), low-income students (Panel C), and low-achieving students (Panel D) in Chicago Public Schools. Panel A shows that a large proportion of classrooms have black students only or non-black students only and a modest proportion of classrooms are racially integrated. Panel B indicates that nearly all classrooms are gender-integrated. Panel C indicates that the majority of classrooms have low-income students only. Panel D shows that the distribution of low-achieving students is closer to a uniform distribution.

interactions between student types (e.g., poor low-achieving, non-poor low-achieving, poor high-achieving, and non-poor high-achieving students), but this is out of scope of this paper and is left for future research. Instead, I look at student gender, SES, and achievement level separately.

Following the steps described in Section 3.4, I make two-year-leave-out forecasts of teacher CA for female students relative to male students, and separately, teacher CA for low-income students relative to high-income students, and teacher CA for low-achieving students relative to high achievers. Panels B, C, and D of Appendix Table A.7 show the quasi-experimental results for gender, socioeconomic status, and achievement level student types, respectively. Panel B shows that female CA has a coefficient of 0.86, implying a forecast bias of 14 percent, but I cannot reject the coefficient is equal to 1. The coefficient of teacher female CA ranges between 0.96–1.11 across grades and subjects, except for middle school math, which is the cause of the large implied bias when all subjects and school levels are pooled. Additionally, I cannot reject that the reference-group VA is forecast unbiased across subsamples.

On the contrary, teacher CA for low-income students fails to pass the quasi-experimental test, but this is driven by the elementary school math and English subsamples, whose coefficient is statistically different from one. For middle school math and English, teacher low-income CA passes the quasi-experimental test, and I cannot reject that its coefficient is equal to one. Additionally, I cannot reject that the reference-group VA is forecast unbiased for middle school levels. One potential reason low-income CA does not pass the quasi-experimental test across some subsamples is the presence of outliers, as can be observed by the coefficients and standard errors that are all closer to zero.

Next, in Panel D, the coefficient of teacher CA for low-achieving students oscillates between 0.50 and 0.77, but I cannot reject it is equal to one at conventional levels in most subsamples. This result indicates a large implied forecast bias in low-income CA and may suggest that a different specification of student types by baseline test scores could better capture disparities in teacher effects, for example, by having more than two student achievement level types or modeling a continuous achievement level type. Nevertheless, the quasi-experimental design developed here could help identify which student characteristics are linked to teachers' disparate causal effects.

8 Sources of heterogeneity and welfare consequences

8.1 What are the sources of heterogeneity in teacher CA?

What kinds of teacher-related policies can be implemented to close achievement gaps may depend on the extent to which teacher CA is malleable or fixed and the sources of heterogeneity in teacher CA. For example, if teacher CA is malleable by teacher behavior in the classroom, then providing professional development may be an avenue to close achievement gaps. On the other hand, if teacher CA is fixed, perhaps due to role model effects, then increasing teacher diversity could exploit the natural variation in teacher effectiveness to close achievement gaps.

I investigate the extent to which teacher CA is malleable and whether teacher characteristics and behaviors, measured by classroom observations and student surveys, explain the variation in teacher CA in Appendix Section C. In short, I find that overall teacher characteristics and the measured teaching practices explain little of the variation in teacher CA. Having more years of experience and a master's degree are positively related with teachers' absolute advantage but negatively related with teacher black CA, which indicates that teachers with experience and graduate degrees increase test scores for everyone but much less for black students. Additionally, while strongly associated with measures of teacher absolute advantage, classroom observations and student surveys are not strongly associated with teacher black CA. Examining the correlation between teacher CA, teaching practices, and classroom learning conditions can provide valuable insights into best practices for closing achievement gaps.[32]

8.2 What are the welfare consequences of considering teacher CA?

I conduct policy counterfactual simulations to estimate the welfare consequences from considering teacher CA into policy decisions. Appendix Section D describes the three sets of counterfactual policy simulations to quantify the potential efficiency and equity gains in student test scores. The simulated policies reflect different ways in which information on teacher black CA could be used. They include (i) teacher accountability policy, which fires the lowest 5 percent effective teachers and replaces them with average teachers; (ii) teacher-to-classroom reassignment policy, which real-

[32]This work is similar in spirit to the literature in economics on the determinants of firm productivity and, more specifically, how managerial practices affect firm productivity (e.g., Bloom and Van Reenen, 2007).

locates teachers to existing classrooms; and (iii) professional development policy, which is intended to equalize disparate impacts for teachers in the lowest 5 percent of the CA distribution. Contrary to the homogeneity assumption in teacher effects, the classroom composition matters under disparate teacher effects as a teacher's total output produced in a classroom (i.e., her total increase in test scores) is a linear combination of her race-specific effects weighted by her classroom's racial composition. I describe here the teacher-to-classroom reassignment policy as this is in theory a resource-neutral policy.

Teacher-to-classroom reassignment policy. A social planner (e.g., school principal) could use information on teacher CA to better match teachers to students to maximize efficiency and equity. Under the homogeneity assumption of teacher effects, reallocating teachers across classrooms is a zero-sum game as the increase in student test scores when a low-VA teacher is replaced by a high-VA one is met by an equal-sized decrease in test scores when the high-VA teacher leaves her classroom. On the other hand, under the heterogeneity assumption of teacher effects, teachers have varying CAs as different teachers are more effective at raising the test scores of different student subgroups. Therefore, reallocating teachers across classrooms becomes a positive-sum game, and better matching teachers to classrooms could increase student test scores overall.

To maximize student output, the optimal policy matches teachers with greater black CA to classrooms with greater proportions of black students, generating positive assortative matching. Within-school teacher reallocation produces small efficiency gains (less than 0.001σ, or 0.6 percent of the benchmark accountability policy). However, across-school reallocation produces larger efficiency gains of 0.003σ per student on average, which represents about 16 percent of the benchmark policy's impact. Reallocating teachers to maximize student achievement would come at the cost of widening the racial achievement gap by 0.038σ, which is 9.2 percent of the black-non-black gap.

The social planner may care about reducing racial achievement gaps. To minimize racial achievement gaps, one way is to assign teachers with greater absolute advantage, i.e., mean VA $= (\mu_{j,black} + \mu_{j,nonblack})/2$, to classrooms with larger proportions of black students. This gap-minimizing allocation policy would reduce the racial achievement gap by 0.150σ, which is 36.2 percent of the black-non-black achievement gap. This equity gain comes with a trade-off of reducing efficiency by -0.005σ because the increase in black students' scores does not fully compensate the reduction in non-black students' scores.

Next, I perform a similar gap-minimizing reallocation exercise subject to not decreasing total students' test scores. This constrained gap-minimizing allocation would reduce the black-non-black gap by 3.6 percent, without reducing average test scores. These results point to the potential of increasing the efficiency and equity of education systems when teacher CA is considered in policy decisions.

It is worth highlighting that these gains are based on the conservative assumption that teacher effects are observed with error. Therefore, improving the accuracy of the teacher CA quality measure could increase these gains even more. It is important to note that the reallocation exercise is based on a reduced vector of student characteristics, meaning that the matching procedure does not account for all potentially relevant student dimensions. This simplification could affect the estimated gains, as they might be higher or lower if multidimensional matching were fully considered. For instance, Ahn et al. (2024), who adopt a multidimensional matching approach, find that optimal teacher allocation across schools within a district results in a 0.05 SD gain in test scores, with greater benefits for black male students, whose gains are approximately 0.07σ.

9 Conclusion

This paper develops a framework to test for the existence of student-teacher match quality. First, I develop a flexible disparate value-added (DVA) model, where a teacher's impact on student test scores depends on student types and fluctuates over time. Second, I propose a novel measure of teacher quality, revealed comparative advantage (CA), which captures a teacher's ability to reduce outcome gaps, such as racial achievement gaps. Third, I validate this measure through a teacher turnover quasi-experiment that exploits the fact that students are exposed to teachers with different comparative advantages while progressing through school. Applying this model to Chicago Public Schools data, quasi-experimental results show that teacher CA is forecast unbiased, with teachers who excel in improving outcomes for black students significantly raising their test scores without negatively affecting non-black students. Policy simulations highlight the potential for significant efficiency and equity gains by incorporating teacher CA into policy decisions, such as teacher-to-classroom reallocation decisions, potentially reducing the black-non-black achievement gap by up to 36 percent (or, by extrapolation, the black-white achievement gap by up to 15 percent).

Obtaining the teacher CA quality measure is readily available and requires the same information used to compute VA measures of teacher quality often used in teacher evaluations. Using the CA quality measure, however, must come with caution as the same criticisms raised for teacher VA may apply to teacher CA (e.g., Rothstein, 2017). In particular, small sample problems are compounded as student-type-specific teacher effects are estimated with a smaller set of students. This problem could be overcome by using multiple years of data or incorporating other measures of teacher quality to improve precision.

Another caution in incorporating teacher CA into policy decisions is that it may induce behavioral responses from teachers that could be detrimental to their students. For example, a teacher with a comparative *disadvantage* for black students may decrease non-black students' test scores rather than increase black students' test scores in order to reduce her disparate impacts. This could be overcome by using teacher mean VA (absolute advantage) or other metrics in addition to teacher CA to capture how teachers affect all students on average. Last, any policy contemplating teacher CA must consider that part of the variation in teacher CA may be malleable (e.g., due to teacher behavior) and another part may be fixed (e.g., due to role model effects).

The results of this paper have implications for the design of educational policies. Often, teacher-related policies assume that the best teacher is the best for everyone (Aucejo et al., 2022). However, the existence of match effects between teachers and students imply that the best teacher for some students may not be the best one for others. As a result, incorporating heterogeneity-based performance measures has the potential to improve the efficiency and equity of education systems.

References

Ahn, T., Aucejo, E., James, J., et al. (2024). The importance of matching effects for labor productivity: Evidence from teacher-student interactions. Technical report, Working Paper.

Ang, D. (2021). The effects of police violence on inner-city students. *The Quarterly Journal of Economics*, 136(1):115–168.

Aucejo, E., Coate, P., Fruehwirth, J. C., Kelly, S., and Mozenter, Z. (2022). Teacher effectiveness and classroom composition: Understanding match effects in the classroom. *The Economic Journal*, 132(648):3047–3064.

Aucejo, E. and James, J. (2021). The path to college education: The role of math and verbal skills. *Journal of Political Economy*, 129(10):2905–2946.

Bacher-Hicks, A., Kane, T. J., and Staiger, D. O. (2014). Validating teacher effect estimates using changes in teacher assignments in los angeles. Technical report, National Bureau of Economic Research.

Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage 1. *The manchester school*, 33(2):99–123.

Barrios-Fernández, A. and Riudavets-Barcons, M. (2024). Teacher value-added and gender gaps in educational outcomes. *Economics of Education Review*, 100:102541.

Bates, M., Dinerstein, M., Johnston, A. C., and Sorkin, I. (2024). Teacher labor market policy and the theory of the second best. *The Quarterly Journal of Economics*, page qjae042.

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

Bloom, N. and Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, 122(4):1351–1408.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–79.

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

Chicago Public Schools (2014). Cps framework for teaching: Companion guide.

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, 84(3):517–544.

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

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, 86(1):195–210.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human resources*, 42(3):528–554.

Dee, T. S. and Penner, E. K. (2017). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal*, 54(1):127–166.

Delgado, W. and Sartain, L. (2023). Context matters: Teacher performance ratings and classroom characteristics. *Unpublished manuscript.*

Delhommer, S. (2022). High school role models and minority college achievement. *Economics of Education Review*, 87:102222.

Egalite, A. J. and Kisida, B. (2018). The effects of teacher match on students' academic perceptions and attitudes. *Educational Evaluation and Policy Analysis*, 40(1):59–81.

Egalite, A. J., Kisida, B., and Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, 45:44 – 52.

Fox, L. (2016). Playing to teachers' strengths: Using multiple measures of teacher effectiveness to improve teacher assignments. *Education Finance and Policy*, 11(1):70–96.

Fryer Jr, R. G. (2011). Racial inequality in the 21st century: The declining significance of discrimination. In *Handbook of labor economics*, volume 4, pages 855–971. Elsevier.

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

Gershenson, S., Holt, S. B., and Papageorge, N. W. (2016). Who believes in me? the effect of student–teacher demographic match on teacher expectations. *Economics of education review*, 52:209–224.

Graham, B. S., Ridder, G., Thiemann, P. M., and Zamarro, G. (2020). Teacher-to-classroom assignment and student achievement. Technical report, National Bureau of Economic Research.

Hanushek, E. A. (2009). *Teacher deselection*, pages 165–180. Urban Institute Press, Washington, DC.

Holder, M. and Aja, A. A. (2021). *Afro-Latinos in the US Economy*. Rowman & Littlefield.

Irvine, J. J. (1989). Beyond role models: An examination of cultural influences on the pedagogical perspectives of black teachers. *Peabody Journal of Education*, 66(4):51–63.

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., Kiguel, S., Porter, S. C., and Easton, J. Q. (2024). Who benefits from attending effective high schools? *Journal of Labor Economics*, 42(3):717–751.

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

Koedel, C., Mihaly, K., and Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47:180–195.

Lefgren, L. and Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis*, 34(1):109–121.

Lindsay, C. A. and Hart, C. M. (2017). Exposure to same-race teachers and student disciplinary outcomes for black students in north carolina. *Educational Evaluation and Policy Analysis*, 39(3):485–510.

Lockwood, J. and McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, 4(4):439–467.

Loeb, S., Soland, J., and Fox, L. (2014). Is a good teacher a good teacher for all? comparing value-added of teachers with their english learners and non-english learners. *Educational Evaluation and Policy Analysis*, 36(4):457–475.

Master, B., Loeb, S., Whitney, C., and Wyckoff, J. (2016). Different skills? identifying differentially effective teachers of english language learners. *The Elementary School Journal*, 117(2):261–284.

Michelmore, K. and Dynarski, S. (2017). The gap within the gap: Using longitudinal data to understand income differences in educational outcomes. *AERA Open*, 3(1):2332858417692958.

Neal, D. A. and Johnson, W. R. (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy*, 104(5):869–895.

Papay, J. P. and Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, 130:105–119.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage.

Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., and Zárate, R. C. (2019a). Gender achievement gaps in us school districts. *American Educational Research Journal*, 56(6):2474–2508.

Reardon, S. F., Kalogrides, D., and Shores, K. (2019b). The geography of racial/ethnic test score gaps. *American Journal of Sociology*, 124(4):1164–1221.

Redding, C. (2019). A teacher like me: A review of the effect of student–teacher racial/ethnic matching on teacher perceptions of students and student academic and behavioral outcomes. *Review of Educational Research*, 89(4):499–535.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American economic review*, 94(2):247–252.

Rothstein, J. (2017). Measuring the impacts of teachers: comment. *American Economic Review*, 107(6):1656–1684.

Staiger, D. O. and Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic perspectives*, 24(3):97–118.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, 100:61–78.

Tables

Table 1: Descriptive Statistics

	Mean (1)	S.D. (2)	N (3)
Panel A: Student characteristics			
Number of subject-school years per student	5.39	[2.96]	331,369
Test scores (SD)	0.11	[0.93]	1,785,027
Female	0.51		1,785,027
Age	10.68	[1.67]	1,785,027
White, non-Hispanic	0.10		1,785,027
Black, non-Hispanic	0.37		1,785,027
Hispanic	0.48		1,785,027
Free or reduced price lunch eligible	0.86		1,785,027
Repeating grade	0.01		1,785,027
Panel B: Classroom characteristics			
Class size (teacher-subject-year cell size)	38.87	[26.83]	48,393
With >= 7 of each race	0.14		48,393
With >= 7 of each gender	0.93		48,393
With >= 7 of each poor status	0.18		48,393
With >= 7 of each achievement lvl	0.70		48,393

Notes: Data come from de-identified administrative data of Chicago Public Schools. Sample is restricted to students in analytic sample used to estimate teacher CA. First column shows the mean, second column the standard deviation, and third column the number of observations. For panel A, the number of observations for the first row is the number of unique students, and for the other rows it is the number of student-subject-year observations. For panel B, the number of observations is the number of unique classrooms.

Table 2: Racial Achievement Gaps in Chicago Public Schools

	Test scores			
	(1)	(2)	(3)	(4)
Black, non-Hispanic	-0.883***	-0.184***	-0.346***	-0.077***
	(0.005)	(0.002)	(0.003)	(0.001)
Hispanic	-0.710***	-0.138***		
	(0.005)	(0.002)		
Other races	0.006	0.032***		
	(0.009)	(0.002)		
Lag score		0.789***		0.804***
		(0.001)		(0.001)
R2	0.101	0.671	0.033	0.668
N	1,785,027	1,785,027	1,785,027	1,785,027

Notes: Table shows OLS regression of test scores (math and English) on race/ethnicity indicators. Sample is restricted to students in analytic sample used to estimate teacher CA. Standard errors are clustered at the student level and reported in parentheses. There is one observation per each student-subject-year across specifications. Column 1 controls for race and ethnicity, Column 2 in addition includes prior test scores. Columns 3 and 4 follow similar specifications but only include black race indicator. The omitted racial group in Columns 1 and 2 is white non-Hispanic students and in Columns 3 and 4 is non-black students. All specifications control for subject-by-school level dummies and year dummies. Asterisks denote *** $p<0.001$, ** $p<0.05$, * $p<0.1$.

Table 3: Variance-Covariance Matrix of Race-Specific Teacher Value-Added

	Math			English		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Elementary schools						
Var 1:	Black VA	non-Black VA	Black VA	Black VA	non-Black VA	Black VA
Var 2:	Black VA	non-Black VA	non-Black VA	Black VA	non-Black VA	non-Black VA
Lag of var 2						
1	0.040	0.048	0.040	0.022	0.014	0.011
	[0.45]	[0.621]	[0.443]	[0.362]	[0.424]	[0.217]
2	0.031	0.044	0.032	0.020	0.013	0.006
	[0.334]	[0.558]	[0.356]	[0.322]	[0.39]	[0.118]
3	0.025	0.041	0.032	0.016	0.010	0.004
	[0.252]	[0.521]	[0.317]	[0.264]	[0.313]	[0.069]
4	0.028	0.035	0.032	0.011	0.008	0.003
	[0.312]	[0.473]	[0.351]	[0.189]	[0.246]	[0.068]
5	0.025	0.033	0.024	0.012	0.006	0.007
	[0.321]	[0.468]	[0.283]	[0.201]	[0.196]	[0.151]
6	0.024	0.034	0.032	0.007	0.008	0.003
	[0.27]	[0.477]	[0.376]	[0.117]	[0.24]	[0.059]
7	0.022	0.030	0.030	0.019	0.004	0.008
	[0.242]	[0.436]	[0.392]	[0.278]	[0.149]	[0.164]
Panel B. Middle schools						
Var 1:	Black VA	non-Black VA	Black VA	Black VA	non-Black VA	Black VA
Var 2:	Black VA	non-Black VA	non-Black VA	Black VA	non-Black VA	non-Black VA
Lag of var 2						
1	0.014	0.017	0.015	0.008	0.007	0.003
	[0.341]	[0.547]	[0.371]	[0.285]	[0.354]	[0.097]
2	0.010	0.013	0.012	0.007	0.006	0.003
	[0.228]	[0.428]	[0.302]	[0.226]	[0.292]	[0.132]
3	0.006	0.013	0.013	0.005	0.005	0.002
	[0.15]	[0.454]	[0.321]	[0.173]	[0.247]	[0.079]
4	0.008	0.012	0.012	0.005	0.005	0.003
	[0.184]	[0.412]	[0.311]	[0.189]	[0.242]	[0.095]
5	0.005	0.012	0.012	0.005	0.003	0.001
	[0.124]	[0.402]	[0.311]	[0.194]	[0.153]	[0.046]
6	0.005	0.012	0.011	0.006	0.004	0.004
	[0.126]	[0.409]	[0.288]	[0.246]	[0.237]	[0.155]
7	0.003	0.012	0.009	0.006	0.003	0.000
	[0.075]	[0.393]	[0.24]	[0.218]	[0.157]	[-0.003]

Panel C: Within-year variance and covariance components for elementary schools

	Black VA	non-Black VA	Black VA	non-Black VA
Total SD	0.356	0.281	0.336	0.237
σ_{ε_k}	0.278	0.213	0.286	0.211
$\sigma^2_{\mu_k} + \sigma^2_{\theta_k}$	0.078	0.068	0.049	0.026
σ_{μ_k} (estimated)	0.231	0.233	0.168	0.132
$\sigma_{\mu_k \mu_m} + \sigma_{\theta_k \theta_m}$	0.022		0.056	
$\sigma_{\mu_k \mu_m}$ (estimated)	0.048		0.022	

Panel D: Within-year variance and covariance components for middle schools

	Black VA	non-Black VA	Black VA	non-Black VA
Total SD	0.259	0.213	0.280	0.234
σ_{ε_k}	0.222	0.184	0.257	0.220
$\sigma^2_{\mu_k} + \sigma^2_{\theta_k}$	0.036	0.029	0.023	0.015
σ_{μ_k} (estimated)	0.142	0.141	0.110	0.089
$\sigma_{\mu_k \mu_m} + \sigma_{\theta_k \theta_m}$	0.010		0.025	
$\sigma_{\mu_k \mu_m}$ (estimated)	0.015		0.003	

Notes: Table presents population parameter estimates of race-specific teacher VA by subject and school level. Columns 1–4 show estimates for math and Columns 5–8 for English. Panels A and B report the variance-covariance matrix and correlation (in brackets) for elementary and middle school levels, respectively. Panels C and D report estimates of the within-year variance and covariance of the parameters for elementary and middle grades, respectively. See Section 6.2 for description of this table.

Table 4: Estimates of Forecast Bias

	Scores in year t								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Black students			non-Black students			Black and non-Black students		
Black CA	0.835		1.090	-2.085		0.017			
	(0.057)		(0.055)	(0.055)		(0.043)			
Matched black CA							0.831		1.097
							(0.057)		(0.052)
Ref. VA		0.914	0.974		1.007	1.011		0.985	0.999
		(0.029)	(0.029)		(0.014)	(0.018)		(0.013)	(0.013)
Ho: CA = 1	0.004	.	0.098	0.000	.	0.000	0.003	.	0.065
Ho: ref. VA = 1	.	0.003	0.365	.	0.616	0.533	.	0.268	0.964
Ho: CA = 0	0.000	.	0.000	0.000	.	0.690	0.000	.	0.000
R2	0.004	0.021	0.027	0.025	0.054	0.054	0.002	0.041	0.043
N	382,535	382,535	382,535	771,005	771,005	771,005	1,153,540	1,153,540	1,153,540

Notes: Table shows OLS regression of residualized student test scores on teacher comparative advantage and reference-group value-added. Standard errors are clustered by school-cohort and reported in parentheses. Sample includes students in analytic sample used to estimate teacher CA and is restricted to those whose teachers have non-missing leave-out teacher CA. There is one observation per each student-subject-year in all regressions. Table splits the sample by student race: Columns 1–3 restrict sample to Black students, Columns 4–6 to non-Black students, and Columns 7–9 pool both Black and non-Black students. Explanatory variables include teacher black CA (Columns 1–6), matched teacher CA (Columns 7–9) and non-black-specific teacher VA (across specifications). Regressions also include year dummies and subject-by-school level dummies. Teacher CA and VA are calculated as described in Section 3.3, and matched teacher CA is the interaction term between teacher CA and black race indicator. P-values of tests that the coefficient of teacher CA is 1, teacher VA is 1, and teacher CA is 0 are reported.

Table 5: Quasi-Experimental Estimates of Forecast Bias

	Δ scores								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Black students			Non-Black students			Black and non-Black students		
Δmean black CA	0.618		1.058	-0.468		-0.072			
	(0.107)		(0.118)	(0.062)		(0.061)			
Δmean matched black CA							0.617		1.039
							(0.107)		(0.107)
Δmean ref. VA		0.466	0.970		0.922	0.886		0.789	0.934
		(0.091)	(0.109)		(0.082)	(0.087)		(0.066)	(0.070)
Ho: CA = 1	0.000	.	0.627	0.000	.	0.000	0.000	.	0.718
Ho: ref. VA = 1	.	0.000	0.785	.	0.341	0.190	.	0.001	0.347
Ho: CA = 0	0.000	.	0.000	0.000	.	0.240	0.000	.	0.000
R2	0.013	0.008	0.033	0.013	0.037	0.038	0.006	0.024	0.035
N	9,575	9,575	9,575	10,094	10,094	10,094	19,669	19,669	19,669

Notes: Table shows OLS regression of changes in average test scores at the student type-school-grade-subject level on changes in estimated teacher CA and VA at the school-grade-subject level. There is one observation per each student type-school-grade-subject in all regressions. Standard errors are clustered by school-cohort and reported in parentheses. Underlying sample includes students in the core sample (described in Section 5.2). Teacher CA and VA are based on student type-specific VA leave-out predictions that excludes years t and $t-1$. Table splits the sample by student race: Columns 1–3 restrict sample to black students, Columns 4–6 to non-black students, and Columns 7–9 pool both black and non-black students. Explanatory variables include changes in teacher black CA (Columns 1–6), matched teacher CA (Columns 7–9) and non-black-specific teacher VA (across specifications). Regressions also include year dummies and subject-by-school level dummies, and weight observations by the number of students in the student type-school-grade cells. Teacher CA and VA are calculated as described in Section 3.3, and matched teacher CA is the interaction term between teacher CA and black race indicator. P-values of tests that the coefficient of changes in teacher CA is 1, teacher VA is 1, and teacher CA is 0 are reported.

Table 6: Robustness Checks: Quasi-Experimental Estimates of Forecast Bias

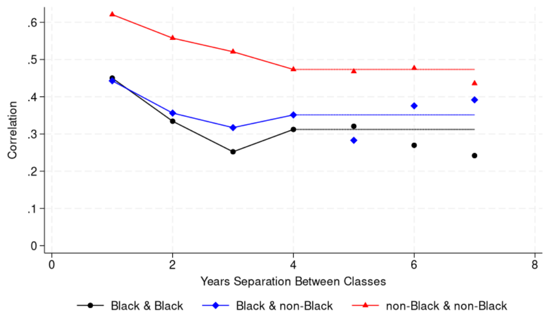
	Δ score	Δ score (past + future data)	Δ score	Δ score	Δ score	Δ other subj. score	Δ other subj. score	Δ lag score	Δ score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Δmean matched black CA	1.039	0.935	1.111	0.561	0.900	0.426	0.222	0.534	1.262
	(0.107)	(0.072)	(0.094)	(0.127)	(0.171)	(0.103)	(0.207)	(0.100)	(0.239)
Δmean ref. VA	0.934	0.909	0.947	0.647	0.943	0.434	0.324	0.448	0.794
	(0.070)	(0.046)	(0.072)	(0.067)	(0.086)	(0.067)	(0.166)	(0.070)	(0.141)
Ho: CA = 1	0.718	0.365	0.239	0.001	0.558	0.000	0.000	0.000	0.272
Ho: ref. VA = 1	0.347	0.049	0.459	0.000	0.511	0.000	0.000	0.000	0.146
Ho: CA = 0	0.000	0.000	0.000	0.000	0.000	0.000	0.283	0.000	0.000
Year fixed effects	X	X				X	X	X	X
School-race year fixed effects			X	X	X				
Lagged score controls				X					
Lag and lead Δ in CA and VA				X	X				
Other-subject Δ in CA and VA						X	X		
OLS or IV	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	IV
Grades	3 to 8	3 to 8	3 to 8	3 to 8	3 to 8	3 to 5	6 to 8	3 to 8	3 to 8
R2	0.035	0.043	0.308	0.446	0.036	0.043	0.021	0.010	0.033
N	19,669	28,484	19,669	7,953	7,953	10,713	4,826	19,669	19,669

Notes: Table shows OLS regression of changes in average test scores at the student type-school-grade-subject level on changes in estimated teacher black CA and VA at the school-grade-subject level. There is one observation per each student type-school-grade-subject in all regressions. Standard errors are clustered by school-cohort and reported in parentheses. Underlying sample includes students in the core sample (described in Section 5.2). Regressions weight observations by the number of students in the student type-school-grade cells. Teacher CA and VA are based on student type-specific VA leave-out predictions that excludes years t and $t-1$. Explanatory variables are changes in matched teacher black CA and changes in non-black-specific teacher VA in all specifications. Column 1 reproduces Column 9 in Panel A of Table 5. Column 2 reproduces results but by using teacher CA and VA forecasts based on prior and future years of data. Column 3 uses forecasts based on prior data only and controls for school-student type-year fixed effects. Column 4 in addition includes leads and lags of changes in teacher CA and VA and cubic polynomial of changes in lagged black- and non-black-specific mean test scores, while Column 5 includes leads and lags of changes in teacher quality but excludes lagged test score controls. Columns 6 and 7 restrict sample to elementary and middle schools, respectively, and have as dependent variable changes in other-subject test scores and further controls for changes in other-subject CA and VA. Column 8 has as dependent variable lagged changes test scores. Column 9 instruments for both changes in mean teacher CA and VA with the fraction of students in the prior cohort taught by teachers who leave the school multiplied by the mean teacher CA and VA, respectively. P-values of tests that the coefficient of changes in teacher CA is 1, teacher VA is 1, and teacher CA is 0 are reported. F-test from the first stage regression in Column 8 is 10.14.

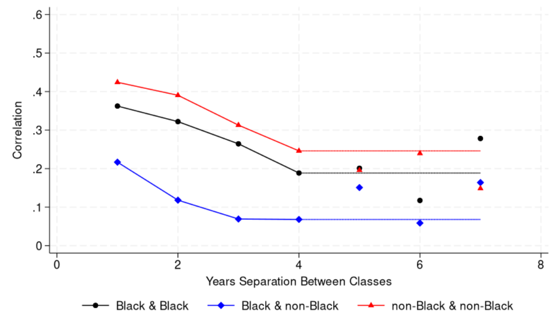
Figures

Figure 1: Autocorrelation and Cross-year Cross-correlation of Race-Specific Teacher Value-Added by Subject and School Level

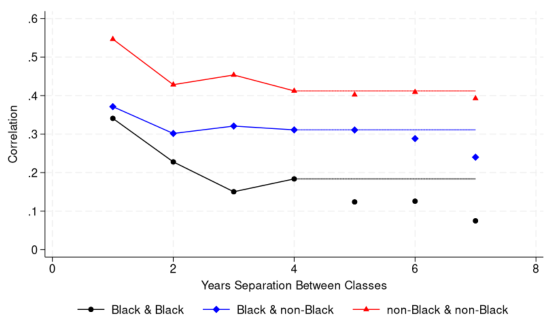
(a) Math, elementary school



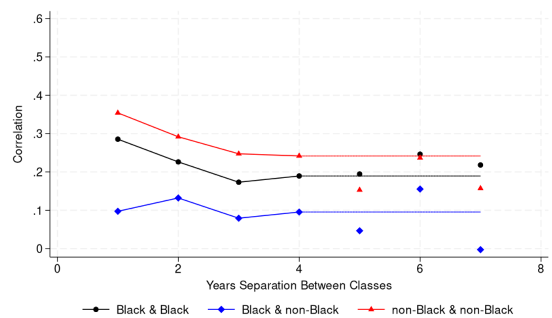
(b) English, elementary school



(c) Math, middle school



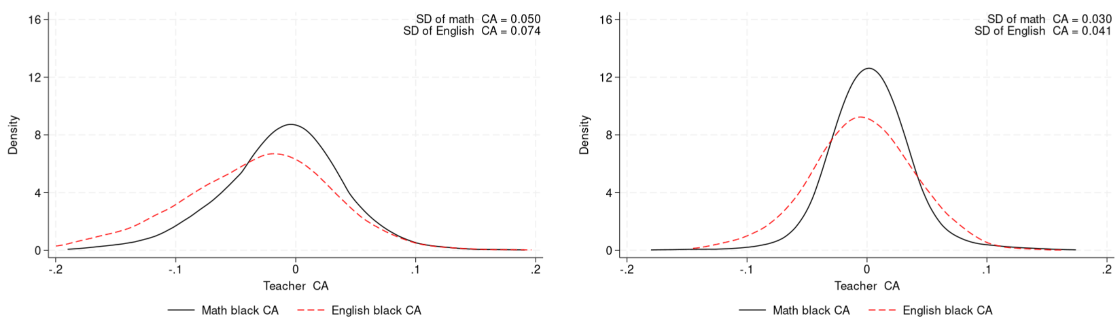
(d) English, middle Schools



Notes: Figures plot the autocorrelation and cross-year crosscorrelation of race-specific teacher value-added by subject and school level. To calculate these values, I first split sample by student type and residualize test scores using within-teacher variation with respect to VA controls. Second, I calculate precision-weighted mean test score residuals across classrooms for each teacher-year, separately by student type. Third, I calculate the autocorrelation coefficients as the correlation across years between mean test score residuals of same student type for a given teacher, weighting by the sum of students from that type taught in the two years. Similarly, I calculate the cross-year crosscorrelation coefficients as the correlation across years between mean test score residuals for different student types for a given teacher, weighting by the sum of students from these types. Panel A shows elementary school math, Panel B elementary school English, Panel C middle school math, and Panel D middle school English. The black line (Black & Black) represents the autocorrelation of black-specific VA; the blue line (Black & non-Black) is the cross-year crosscorrelation of black- and non-black-specific VA; and the red line (non-Black & non-Black) is the autocorrelation of non-black-specific VA.

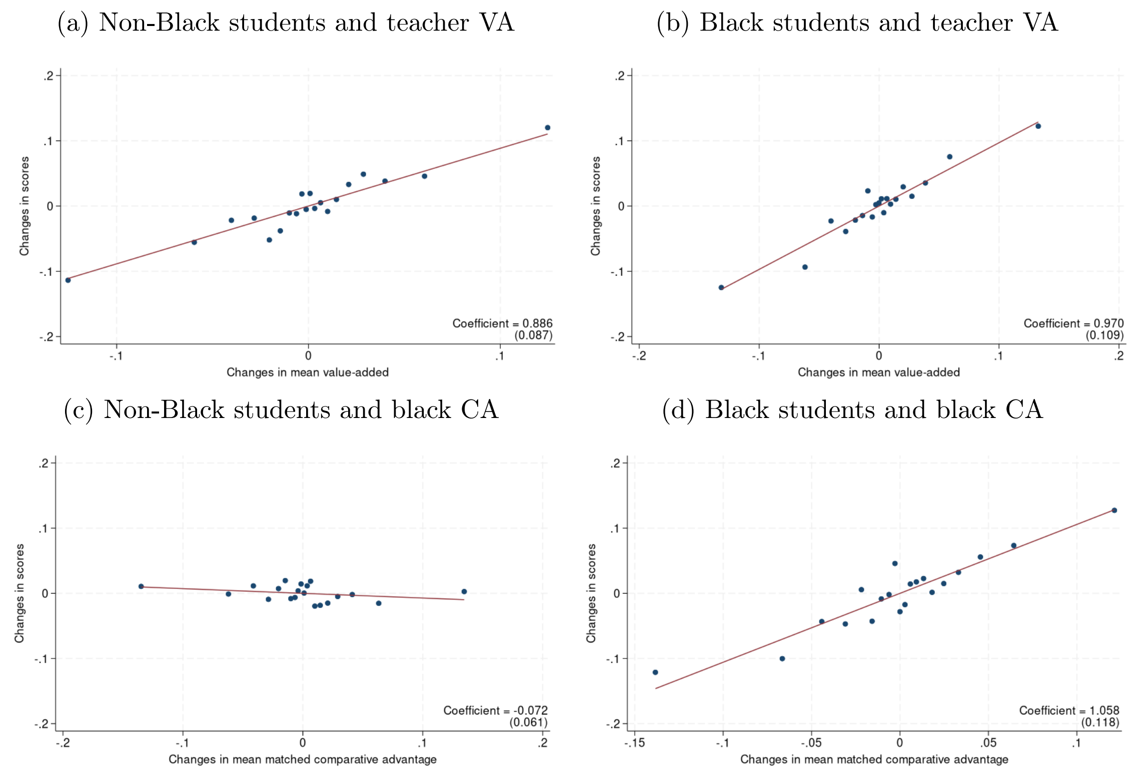
Figure 2: Empirical Distribution of Teacher Black Revealed Comparative Advantage by Subject and School Level

(a) Elementary schools (b) Middle schools



Notes: Figures plot kernel densities of teacher black comparative advantage estimates by subject and school level. Panel A shows the densities for elementary school level and Panel B for middle school level. Standard deviations of the empirical distributions are reported in each figure. The densities use a bandwidth of 0.02. and are weighted by the total number of student test score observations in the classroom used to construct student-type specific VA estimates.

Figure 3: Effects of Changes in Teaching Staff on Black and non-Black Students' Test Scores



Notes: Figures plot changes in average test scores across cohorts versus changes in teacher black comparative advantage and reference-group value-added. Each figure is a binned scatterplot, where the sample is grouped into 20 equal-size groups based on values of the x-axis variable and the mean of the y-axis variable is plotted for each group. The red line is the best linear fit based on the underlying data. Panels A and B plot changes in non-black and black students' test scores, respectively, versus changes in teacher VA, and Panels C and D plot these changes versus changes in teacher black CA. Panels A and C correspond to Column 3 of Table 5, and Panels B and D correspond to Column 6.

Appendix (for online publication)

A Appendix Tables and Figures

Appendix Tables

Table A.1: Descriptive Statistics of Teachers

	Mean (1)	S.D. (2)	N (3)
Number of subject-school years per teacher	3.38	[2.24]	9,740
Number of school years per teacher-subject	3.04	[2.09]	15,894
Female	0.85		47,258
White, non-Hispanic	0.46		46,901
Black, non-Hispanic	0.29		46,901
Hispanic	0.20		46,901
Age	40.83	[10.59]	47,358
Experience	9.57	[7.91]	30,721
Tenured	0.50		47,369
Master's degree	0.69		46,651
Ever taught >= 7 of each race	0.18		48,393
Ever taught >= 7 of each gender	0.95		48,393
Ever taught >= 7 of each poor status	0.23		48,393
Ever taught >= 7 of each achievement lvl	0.79		48,393

Notes: Data come from de-identified administrative data of Chicago Public Schools. Sample is restricted to teachers of students in analytic sample used to estimate teacher CA. First column shows the mean, second column the standard deviation, and third column the number of observations. The number of observations for the first row is the number of unique teachers, for the second row it is the number of unique teacher-subject cells, and for the other rows it is the number of teacher-subject-year observations. A teacher is said to have ever taught at least seven black and seven non-black students if any of her classrooms in the present or the past have had at least seven black and seven non-black students. This is similarly defined for ever teaching to both genders, poverty status, and achievement level student types.

Table A.2: Racial Achievement Gaps in Chicago Public Schools by Subject and School Level

	Test scores							
	Elementary schools		Middle schools		Elementary schools		Middle schools	
	Math (1)	English (2)	Math (3)	English (4)	Math (5)	English (6)	Math (7)	English (8)
Black, non-Hispanic	-0.968***	-0.851***	-0.925***	-0.785***	-0.414***	-0.276***	-0.411***	-0.282***
	(0.007)	(0.006)	(0.007)	(0.007)	(0.004)	(0.004)	(0.004)	(0.004)
Hispanic	-0.746***	-0.757***	-0.685***	-0.650***				
	(0.007)	(0.006)	(0.007)	(0.006)				
Other races	0.067***	-0.095***	0.125***	-0.074***				
	(0.011)	(0.010)	(0.012)	(0.011)				
R2	0.120	0.096	0.113	0.080	0.043	0.021	0.045	0.024
N	439,098	429,779	459,550	456,600	439,098	429,779	459,550	456,600

Notes: Table shows OLS regression of test scores (math or English) on race/ethnicity indicators, separately by subject and school level. Sample is restricted to students in analytic sample used to estimate teacher CA. Standard errors are clustered at the student level and reported in parentheses. There is one observation per each student-subject-year across specifications. The omitted racial group is white non-Hispanic students for Columns 1–4 and non-black students for Columns 5–8. All specifications control for subject-by-school level dummies and year dummies. Asterisks denote *** $p<0.001$, ** $p<0.05$, * $p<0.1$.

Table A.3: Number of Observations Used to Estimate the Variance-Covariance Matrix of Race-Specific Teacher Value-Added

	Math			English		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Elementary schools						
Var 1:	Black VA	non-Black VA	Black VA	Black VA	non-Black VA	Black VA
Var 2:	Black VA	non-Black VA	non-Black VA	Black VA	non-Black VA	non-Black VA
Lag of var 2						
1	6262	12360	1588	6348	13068	1575
2	3996	8300	1056	3996	8850	1052
3	2510	5522	711	2542	5904	701
4	1526	3490	454	1502	3710	440
5	872	1886	247	814	2022	241
6	500	1064	148	482	1156	150
7	182	444	69	200	494	71
Panel B. Middle schools						
Var 1:	Black VA	non-Black VA	Black VA	Black VA	non-Black VA	Black VA
Var 2:	Black VA	non-Black VA	non-Black VA	Black VA	non-Black VA	non-Black VA
Lag of var 2						
1	4264	5534	1974	4466	6780	1843
2	2928	3984	1415	2840	4616	1234
3	2134	2966	1095	1994	3274	896
4	1478	2104	772	1290	2224	608
5	938	1370	513	784	1406	393
6	614	896	366	514	910	271
7	256	378	157	236	408	120

Notes: Table shows the number of observations used to estimate the variance-covariance matrix of race-specific teacher VA presented in Table 3 by subject for elementary schools (Panel A) and middle schools (Panel B). Columns 1–3 show math and Columns 4–6 English.

Table A.4: Estimates of Population Variance and Covariance of Race-Specific Teacher Effects

	Elementary schools		Middle schools	
	Math	English	Math	English
	(1)	(2)	(3)	(4)
$\sigma_{\mu_{Black}}$	0.231	0.168	0.142	0.110
$\sigma_{\mu_{nonBlack}}$	0.233	0.132	0.141	0.089
$\rho_{\mu_{nonBlack}\mu_{Black}}$	0.890	0.993	0.764	0.326
σ_{CA}	0.109	0.041	0.097	0.117
... perc. BnB gap	41.4%	14.7%	23.7%	41.4%
Naive $Corr(VA_{nonBlack} VA_{Black})$	0.973	0.534	0.954	0.563
Naive $SD(BlackCA)$	0.050	0.074	0.030	0.041
Stability $Corr(BlackCA_t, BlackCA_{t-1})$	0.712	0.888	0.725	0.607

Source: Table reports estimates of the population variance and correlation of race-specific teacher VAs and variance of teacher black CA by subject and school level. Table also expresses the standard deviation of teacher black CA as a percentage of the black-non-black achievement gaps.

Table A.5: Relationship between Measures of Teacher Absolute and Comparative Advantage with Homogeneous Teacher Value-Added

	Mean VA (1)	Ref. group VA (2)	Black CA (3)
Homogeneous VA	0.965 (0.014)	1.040 (0.012)	-0.149 (0.011)
Ho: hom. VA = 1	0.012	0.001	0.000
Ho: hom. VA = 0	0.000	0.000	0.000
R2	0.918	0.895	0.156
N	6,390	6,390	6,390

Notes: Table shows OLS regression of teacher mean VA, reference-group VA, and black CA on homogeneous teacher VA estimates. Standard errors are clustered at the teacher level and reported in parentheses. Sample is restricted to teachers who are ever observed teaching students from both races. There is one observation per each teacher-subject-year across specifications. The dependent variable in Column 1 is the mean of race-specific teacher VAs, in Column 2 is non-black-specific teacher VA, and in Column 3 is teacher black CA. The explanatory variable is homogeneous teacher VA, calculated under the scenario that there is a single student type. Regressions also control for subject-by-school level indicators. P-values of coefficient tests in which null hypothesis is that homogeneous teacher VA is 1 and 0 are reported.

Table A.6: Reproducing Quasi-Experimental Results of Forecast Bias under Homogeneous Teacher Value-Added

	Δ score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Δmean homog. VA	0.984	1.019	0.726	1.061	0.841	1.015	0.877
	(0.070)	(0.069)	(0.063)	(0.109)	(0.070)	(0.158)	(0.127)
Δmean other subject VA					0.351	0.271	
					(0.064)	(0.124)	
Ho: homog. VA = 1	0.818	0.779	0.000	0.574	0.024	0.925	0.333
Ho: other subj. VA = 1					0.000	0.000	
Year f.e.	X				X	X	X
School-year f.e.		X	X	X			
Lagged score controls			X				
Lag and lead Δ in VA			X	X			
OLS or IV	OLS	OLS	OLS	OLS	OLS	OLS	IV
Grades	3 to 8	3 to 8	3 to 8	3 to 8	3 to 5	6 to 8	3 to 8
R2	0.041	0.307	0.712	0.359	0.053	0.035	0.040
N	12,665	12,665	5,106	5,106	7,084	2,992	12,665

Notes: Table reproduces main results of Chetty et al. (2014a), who estimate homogeneous teacher value-added. Columns 1–5 reproduces their Table 4 and Column 6 reproduces their Table 5 Column 1. Results show OLS regression of changes in average test scores at the school-grade-subject level on changes in estimated homogeneous teacher VA at the school-grade-subject level. There is one observation per each school-grade-subject in all regressions. Standard errors are clustered by school-cohort and reported in parentheses. Underlying sample includes students in main sample (described in Section 5.2). Explanatory variable is the change in forecasted homogeneous VA using prior year data, and forecasts excludes years t and $t-1$. Regressions weight observations by the total number of students in the school-grade cell. Column 1 controls for year fixed effects. Column 2 controls for school-year fixed effects. Column 3 in addition includes leads and lags of changes in mean teacher VA and cubic polynomial of change in lagged mean test scores. Columns 4 and 5 split sample by school level and control for year fixed effects and changes in mean other-subject homogeneous VA. Column 6 shows estimates from a 2SLS regression, where teacher homogeneous VA is instrumented by the fraction of students in the prior cohort taught by teachers who exit the school multiplied by the mean homogeneous VA of those teachers.

Table A.7: Quasi-Experimental Results of Forecast Bias for Various Student Types by Subject and School Level

	All	Elementary		Middle	
		Math	English	Math	English
	(1)	(2)	(3)	(4)	(5)
Panel A: Black and non-Black students					
Δmean match black CA	1.039	1.035	1.038	0.965	1.042
	(0.107)	(0.127)	(0.184)	(0.250)	(0.301)
Δmean ref. VA	0.934	0.895	0.992	0.913	1.282
	(0.070)	(0.081)	(0.152)	(0.155)	(0.298)
Ho: CA = 1	0.718	0.781	0.837	0.889	0.890
Ho: ref. VA = 1	0.347	0.197	0.960	0.573	0.346
R2	0.035	0.051	0.028	0.035	0.023
N	19,669	5,958	6,162	3,581	3,959
Panel B: Girls and boys					
Δmean match female CA	0.860	1.012	1.110	0.248	0.963
	(0.204)	(0.275)	(0.350)	(0.598)	(0.448)
Δmean ref. VA	0.974	0.933	1.010	0.950	1.345
	(0.068)	(0.079)	(0.124)	(0.161)	(0.281)
Ho: CA = 1	0.494	0.966	0.753	0.208	0.934
Ho: ref. VA = 1	0.703	0.396	0.933	0.758	0.220
R2	0.032	0.046	0.026	0.031	0.023
N	24,965	7,696	7,953	4,385	4,919
Panel C: Low- and high-income students					
Δmean match low-income CA	0.000	0.225	-0.001	0.920	0.730
	(0.001)	(0.055)	(0.000)	(0.177)	(0.330)
Δmean ref. VA	0.001	0.102	0.000	0.968	1.502
	(0.001)	(0.037)	(0.000)	(0.166)	(0.332)
Ho: CA = 1	0.000	0.000	0.000	0.652	0.412
Ho: ref. VA = 1	0.000	0.000	0.000	0.848	0.131
R2	0.001	0.008	0.004	0.036	0.024
N	21,373	6,412	6,653	3,938	4,360
Panel D: Lower- and higher-achieving students					
Δmean match low-achieving CA	0.724	0.754	0.766	0.565	0.490
	(0.087)	(0.118)	(0.136)	(0.252)	(0.363)
Δmean ref. VA	0.745	0.698	0.680	0.867	0.976
	(0.054)	(0.066)	(0.097)	(0.128)	(0.220)
Ho: CA = 1	0.001	0.037	0.084	0.085	0.160
Ho: ref. VA = 1	0.000	0.000	0.001	0.296	0.914
R2	0.026	0.035	0.020	0.037	0.014
N	24,797	7,674	7,926	4,335	4,853

Notes: Table shows quasi-experimental results for various student types by subject and school level. Column 1 of Panels A and B reproduce results in Table 5, Column 9 of Panels A and B. Panel C performs a similar exercise but instead uses teacher low-income CA, which is constructed as the difference between low- and high-income-specific VAs. Panel D uses teacher low-achieving CA, which is the difference between low- and high-achievement-specific VAs (see Section 7.4 for description of this table). Columns 2–5 split the sample by subject and school level.

Table A.8: Relationship between Teacher Black Comparative Advantage and Teacher Characteristics

	Mean VA > 0	Black CA > 0	Mean VA > 0, Black CA > 0
	(1)	(2)	(3)
Female	0.013	-0.018	-0.006
	(0.028)	(0.023)	(0.018)
Black, non-Hispanic	-0.002	0.034	0.050***
	(0.026)	(0.022)	(0.018)
Hispanic	-0.002	0.007	-0.020
	(0.030)	(0.027)	(0.018)
Other race	0.021	-0.023	0.018
	(0.043)	(0.031)	(0.025)
Age	-0.002**	0.001	0.000
	(0.001)	(0.001)	(0.001)
Years of experience	0.013**	-0.009**	0.006*
	(0.005)	(0.004)	(0.003)
Experience squared	-0.000	0.000*	-0.000
	(0.000)	(0.000)	(0.000)
Tenured	-0.029	0.064**	0.018
	(0.030)	(0.028)	(0.021)
Master's degree	0.064***	-0.045**	-0.011
	(0.024)	(0.020)	(0.015)
R2	0.01	0.04	0.02
N	6,354	6,354	6,354

Notes: Table shows OLS regression of teacher black comparative advantage and mean value-added on teacher characteristics. Standard errors are clustered at the teacher level and reported in parentheses. Sample is restricted to teachers who are ever observed teaching students from both races. There is one observation per each teacher-subject-year across specifications. Dependent variable in Column 1 is whether teacher mean VA is above that of the average teacher (high-effective), Column 2 is whether teacher black CA is above the average (high-equity), and Column 3 is whether both teacher mean VA and black CA are above those of the average teacher's (high-effective and high-equity). Mean VA is the average of race-specific VAs. In addition to the independent variables listed in the table, regressions also include subject-by-school level dummies and year dummies. Asterisks denote *** p<0.001, ** p<0.05, * p<0.1.

Table A.9: Relationship between Teacher Black Comparative Advantage and Classroom Observation Ratings

	Planning and preparation		Classroom environment		Instruction		Professional responsibilities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black CA	0.022	0.030	0.024	0.022	0.032	0.032	0.004	0.006
	(0.064)	(0.059)	(0.060)	(0.055)	(0.060)	(0.054)	(0.057)	(0.053)
Ref. VA	0.282***	0.248***	0.307***	0.265***	0.306***	0.261***	0.230***	0.197***
	(0.041)	(0.038)	(0.039)	(0.036)	(0.042)	(0.038)	(0.037)	(0.034)
Teacher controls		X		X		X		X
R2	0.04	0.12	0.05	0.12	0.05	0.13	0.03	0.10
N	3,609	3,602	3,689	3,681	3,689	3,681	3,616	3,608

Notes: Table shows OLS regression of classroom observation ratings on teacher black comparative advantage and reference-group value-added. Standard errors are clustered at the teacher level and reported in parentheses. Sample is restricted to teachers who are ever observed teaching students from both races and with classroom observation data. There is one observation per each teacher-subject-year across specifications. Observers rated teachers on various teaching practices multiple times in a year and gave scores from 1 to 4, where 1 indicates unsatisfactory and 4 distinguished. Dependent variables are averages of these ratings grouped into four domains: planning and preparation (Columns 1 and 2), classroom environment (Columns 3 and 4), instruction (Columns 5 and 6), and professional responsibilities (Columns 7 and 8). Each dependent variable is standardized to have mean zero and standard deviation one by year-subject-school level. The explanatory variables teacher black CA and VA are rescaled by the estimates of their true variances. All regressions control for subject-by-school level dummies and year dummies. Specifications with teacher controls further include teacher's gender, race and ethnicity, age, years of experience and its squared, tenure status, and master's degree indicator. Asterisks denote *** $p<0.001$, ** $p<0.05$, * $p<0.1$.

Table A.10: Relationship between Teacher Black Comparative Advantage and Student Survey Ratings

	Peer support	Classroom rigor	Academic press	Course clarity	Academic engagement	Academic personalism	Classroom disruptions
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Without teacher controls							
Black CA	0.039	-0.003	0.041	0.028	0.006	0.005	0.018
	(0.044)	(0.048)	(0.048)	(0.045)	(0.056)	(0.057)	(0.059)
Ref. VA	0.104***	0.133***	0.175***	0.137***	0.122***	0.097***	0.147***
	(0.026)	(0.031)	(0.029)	(0.027)	(0.029)	(0.034)	(0.033)
Teacher controls							
R2	0.02	0.02	0.04	0.03	0.02	0.02	0.02
N	3,255	3,255	3,254	3,248	3,249	3,250	3,248
Panel B: With teacher controls							
Black CA	0.048	0.005	0.041	0.036	0.015	0.023	0.026
	(0.044)	(0.046)	(0.047)	(0.044)	(0.056)	(0.056)	(0.056)
Ref. VA	0.098***	0.121***	0.163***	0.129***	0.120***	0.097***	0.127***
	(0.026)	(0.030)	(0.029)	(0.027)	(0.029)	(0.033)	(0.033)
Teacher controls	X	X	X	X	X	X	X
R2	0.02	0.04	0.06	0.04	0.04	0.04	0.06
N	3,236	3,236	3,235	3,229	3,230	3,231	3,229

Notes: Table shows OLS regression of student survey indexes on teacher black comparative advantage and reference-group value-added. Standard errors are clustered at the teacher level and reported in parentheses. Sample is restricted to teachers who are ever observed teaching students from both races and who were linked to student survey data. There is one observation per each teacher-subject-year across specifications. Survey data come from 5Essentials student survey. Dependent variables are classroom-level averages of student-level indexes, where each index is the average across items (see Appendix Table A.2) and are normalized to have mean zero and standard deviation one by year-subject-school level. The explanatory variables teacher black CA and VA are rescaled by the estimates of their true variance. All regressions include subject-by-school level dummies. Specifications with teacher controls (Panel B) further include teacher's gender, race and ethnicity, age, years of experience and its squared, tenure status, and master's degree indicator. Asterisks denote *** $p<0.001$, ** $p<0.05$, * $p<0.1$.

Table A.11: Estimates of Student-Teacher Race and Gender Match Effects Using Within-Student Variation

	Scores	
	(1)	(2)
Race match	0.009**	
	(0.004)	
Gender match	0.005***	
	(0.001)	
Black race match		-0.009*
		(0.005)
Non-Black race match		0.026***
		(0.006)
Female gender match		0.012***
		(0.004)
Male gender match		-0.001
		(0.004)
Teacher controls	X	X
Student f.e.	X	X
R-squared	0.24	0.24
N	1,707,231	1,707,231

Notes: Table reports estimates of student-teacher racial and gender match effects in Chicago Public Schools. These estimates come from student-fixed effect regressions of (residualized) student test scores on indicators whether the student and her teacher share the same race (black or non-black) or same gender (female or male). Residualized test scores come from regressing test scores on VA controls used to estimate race-specific teacher VA and teacher fixed effects. Column 1 controls for race and gender match indicators, and Column 2 disaggregate these indicators into black and non-black racial match and female and male gender match indicators. Specifications control for subject-by-school level indicators and year dummies and cluster standard errors at the cohort level. Asterisks denote *** p<0.001, ** p<0.05, * p<0.1.

Table A.12: Potential Efficiency and Equity Gains from Various Counterfactual Education Policies

Policy	$\Delta Q_{5\%}$	ΔQ			ΔA_0	ΔA_1	$\Delta A_0 - \Delta A_1$	
	σ	σ	% SD meanVA	% bench.	σ	σ	σ	% BnB gap
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Teacher accountability								
True homog VA	0.506	0.025	5.6%	138.8%	0.025	0.025	0.000	0.1%
Benchmark: Homog VA	0.365	0.018	4.0%	100.0%	0.019	0.016	0.004	0.9%
VA + pCA	0.369	0.018	4.1%	101.2%	0.020	0.016	0.004	0.9%
2. Teacher reallocation								
2.1. Maximize output								
Within school	-	0.000	0.0%	0.6%	0.000	0.000	0.001	0.2%
Across schools	-	0.003	0.7%	16.2%	0.014	-0.024	0.038	9.2%
2.2. Minimize gap								
Unconstrained	-	-0.005	-1.2%	-28.7%	-0.050	0.100	-0.150	-36.2%
Subject to $\Delta Q \geq 0$	-	0.001	0.2%	4.0%	-0.004	0.011	-0.015	-3.6%
3. Professional development								
CA	0.038	0.002	0.4%	10.4%	0.000	0.006	-0.006	-1.5%

Notes: Table quantifies efficiency and equity gains from various counterfactual education policies. Gains are expressed in test score standard deviation units per student. The first set of policies is a teacher accountability policy that replaces teachers in the bottom 5 percent of the homogeneous teacher VA distribution with teachers of average quality (rows 1 and 1), and a similar accountability policy that instead replaces teachers in the bottom of the expected impacts in a randomly chosen classroom (row 2). The second set of policies mimics a professional development policy that selects teachers in the bottom 5 percent of the black CA distribution and changes their CA to that of the average teacher (row 4). The following set of policies maximizes total output by reallocating teachers to classrooms within schools (row 5) and across schools (row 6). The last set of policies aims to minimize the racial achievement gap by reallocating teachers to classrooms without any restrictions (row 7) and subject to to not decreasing total test scores (row 8). Simulation in row 1 assumes that the true race-specific teacher VAs are observed, while the rest assumes that teacher VAs are estimating by introducing noise. Column 1 shows the total gains for the 5 percent of classrooms affected by the accountability policies. Column 2 shows the total gains for everyone, Column 3 expresses these gains as percentage of standard deviation units of mean teacher VA, and Column 4 as percentage of the benchmark policy (shown in row 2). Columns 5 and 6 express these gains as average test score gains for non-black and black students, respectively. Column 7 shows changes in the racial achievement gap, where a positive value means an increase and a negative value means a decrease in the gap. Column 9 expresses these equity gains as percentage of the black-non-black achievement gap. Policy simulations use information from elementary school math across all years.

Appendix Figures

Figure A.1: Example of a Rubric Associated with the Classroom Environment Domain

Component	*Unsatisfactory*	*Basic*	*Proficient*	*Distinguished*
2a: Creating an Environment of Respect and Rapport • *Teacher Interactions with Students* • *Student Interactions with Other Students*	Patterns of classroom interactions, both between the teacher and students and among students, are mostly negative and disrespectful. Interactions are insensitive and/or inappropriate to the ages and development of the students, and the context of the class. The net result of interactions has a negative impact on students emotionally and/or academically.	Patterns of classroom interactions, both between the teacher and students and among students, are generally respectful but may reflect occasional inconsistencies or incidences of disrespect. Some interactions are sensitive and/or appropriate to the ages and development of the students, and the context of the class. The net result of the interactions has a neutral impact on students emotionally and/or academically.	Patterns of classroom interactions, both between the teacher and students and among students, are friendly and demonstrate caring and respect. Interactions among students are generally polite and respectful. Interactions are sensitive and appropriate to the ages and development of the students, and to the context of the class. The net result of the interactions has a positive impact on students emotionally and academically.	Patterns of classroom interactions, both between the teacher and students and among students, are highly respectful, reflecting genuine warmth and caring. Students contribute to high levels of civility among all members of the class. Interactions are sensitive to students as individuals, appropriate to the ages and development of individual students, and to the context of the class. The net result of interactions is that of academic and personal connections among students and adults.

Source: Chicago Public Schools (2014).
Notes: Figure shows rubric used to evaluate Component 2a: Creating an environment of respect and rapport, which is part of Domain 2: Classroom environment.

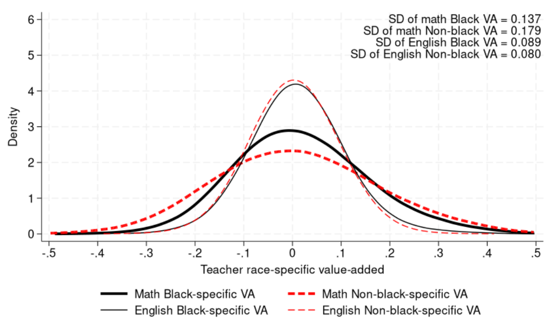
Figure A.2: Survey Questions and Indexes

Survey items	Index
A. How many students in your class. . . 1 Feel it is important to come to school every day? 2 Feel it is important to pay attention in class? 3 Think doing homework is important? 4 Try hard to get good grades?	Peer support
B. How much do you disagree or agree with the following statements about your teacher in your class? My teacher. . . 1 Often connects what I am learning to life outside of the classroom. 2 Encourages students to share their ideas about things we are studying in class. 3 Often requires me to explain my answers. 4 Encourages us to consider different solutions or points of view. 5 Doesn't let students give up when the work gets hard.	Classroom rigor
C. How often does the following occur? In my [TARGET] class, we talk about different solutions or points of view.	
D. How much do you disagree or agree with the following statements about your class? 1 This class really makes me think. 2 I'm really learning a lot in this class.	Academic press
E. To what extent do you disagree or agree with the following statements? In my class, my teacher. . . 1 Expects everyone to work hard. 2 Expects me to do my best all the time. 3 Wants us to become better thinkers, not just memorize things.	
F. In your class, how often. . . 1 Are you challenged? 2 Do you have to work hard to do well? 3 Does the teacher ask difficult questions on tests? 4 Does the teacher ask difficult questions in class?	
G. How much do you disagree or agree with the following statements about your class? 1 I learn a lot from feedback on my work. 2 It's clear to me what I need to do to get a good grade. 3 The work we do in class is good preparation for the test. 4 The homework assignments help me to learn the course material. 5 I know what my teacher wants me to learn in this class.	Course clarity
H. How much do you disagree or agree with the following statements about your class? 1 I usually look forward to this class. 2 I work hard to do my best in this class. 3 Sometimes I get so interested in my work I don't want to stop. 4 The topics we are studying are interesting and challenging.	Academic engagement
I. How much do you disagree or agree with the following statements about your class? The teacher for this class. . . 1 Helps me catch up if I am behind. 2 Is willing to give extra help on schoolwork if I need it. 3 Notices if I have trouble learning something. 4 Gives me specific suggestions about how I can improve my work in this class. 5 Explains things in a different way if I don't understand something in class.	Academic personalisms
J. How much do you disagree or agree with the following statement about your class? 1 I get distracted from my work by other students acting out in this class. 2 This class is out of control. 3 My classmates do not behave the way my teacher wants them to.	Classroom disruptions

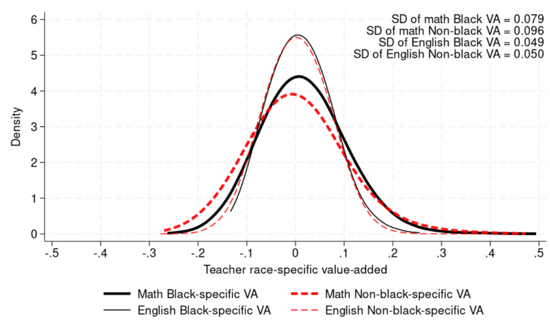
Source: Delgado and Sartain (2023)
Notes: Table shows survey questions from 5Essentials student survey that were consistently asked across years and were about the math and English teachers. These survey questions are grouped into seven categories shown in the second column. Students gave the following responses (with values in parentheses) as follows. For question A: (1) None (2) A few (3) About half (4) Most (5) All. For questions B, D, E, G, H, I, J: (1) Strongly disagree (2) Disagree (3) Agree (4) Strongly agree. For question C: (1) Very little (2) Some (3) Quite a bit (4) A great deal. And for question F: (1) Never (2) Once in a while (3) Most of the time (4) All the time. Peer support: reveals whether prevailing norms among students support academic work.

Figure A.3: Empirical Distribution of Race-Specific Teacher Value-Added Estimates by Subject and School Level

(a) Elementary schools



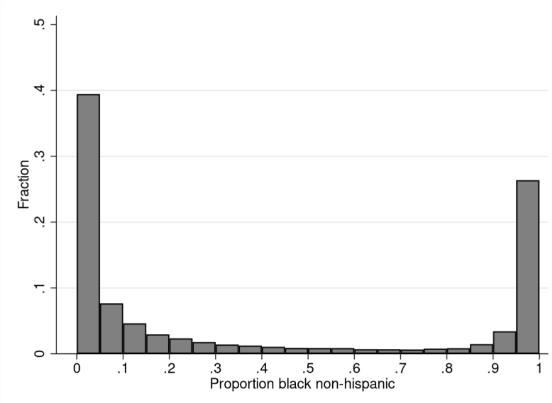
(b) Middle schools



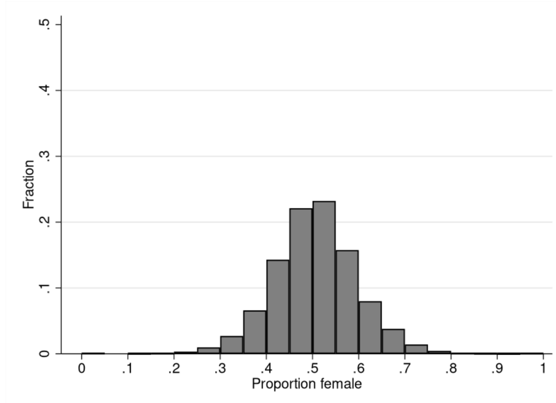
Notes: Figures plot kernel densities of race-specific teacher value-added estimates by subject and school level. Panel A shows the densities for elementary school level and Panel B for middle school level. Standard deviations of the empirical distributions are reported in each figure. The densities use a bandwidth of 0.05. and are weighted by the number of student test score observations belonging to the corresponding student type used to construct student-type specific teacher VA.

Figure A.4: Racial, Gender, Income, and Achievement Level Classroom Composition

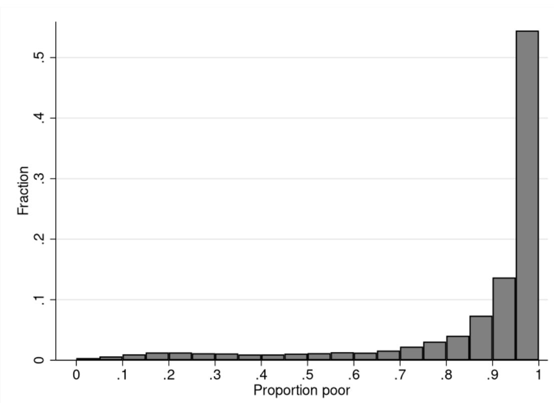
(a) Proportion of black students



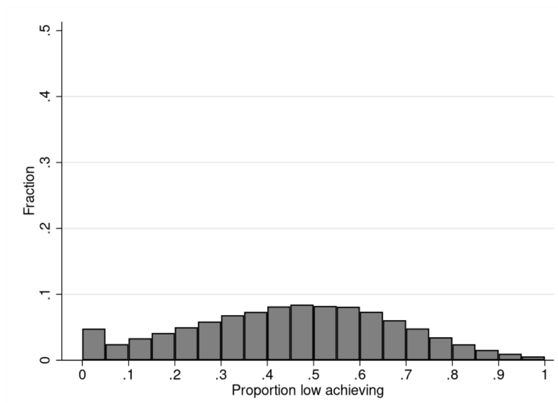
(b) Proportion of female students



(c) Proportion of free/reduced-price lunch students



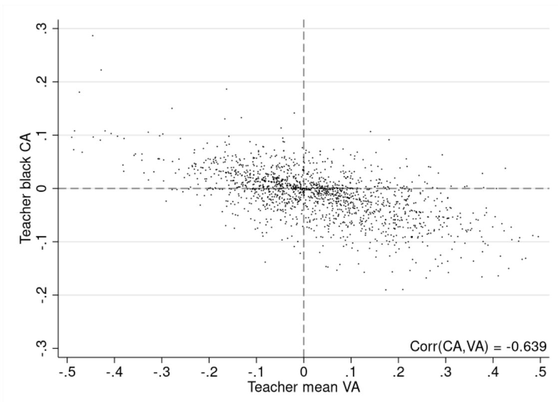
(d) Proportion of lower-achieving students



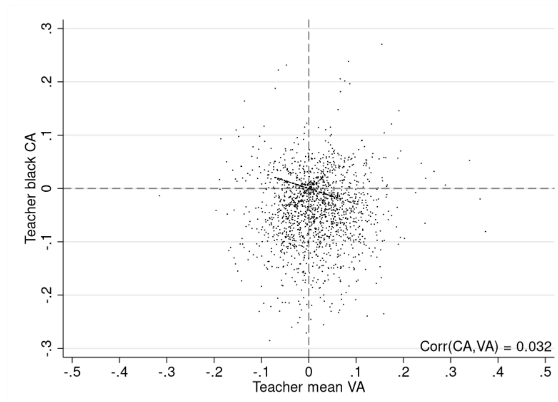
Notes: Figures plot the distribution of the proportion of black students (PanelA), girls (Panel B), free/reduced-price lunch status students (Panel C) and lower-achieving students, defined as those whose baseline test scores rank them below median (Panel D) in Chicago Public Schools classrooms. Each bin is 10 percentage-points wide.

Figure A.5: Relationship between Teacher Black Revealed Comparative Advantage and Mean Value-Added by Subject and School Level

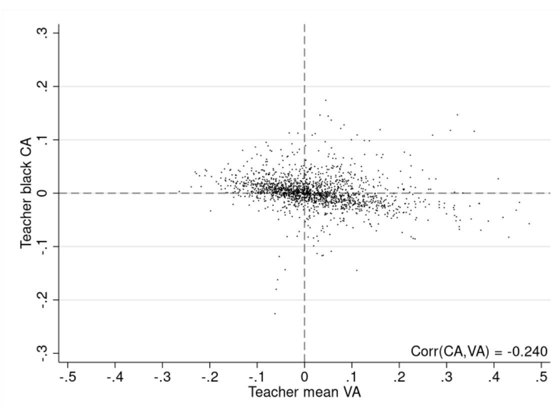
(a) Math, elementary school



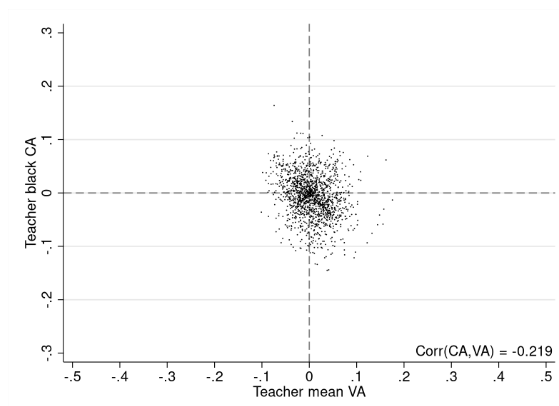
(b) English, elementary school



(c) Math, middle school



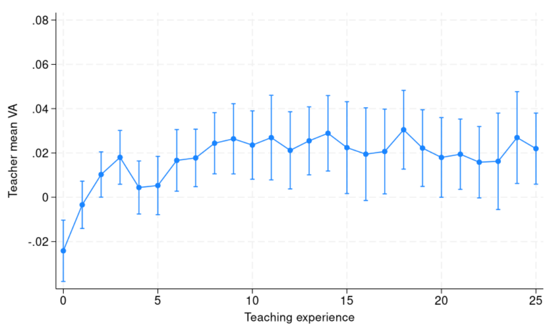
(d) English, middle school



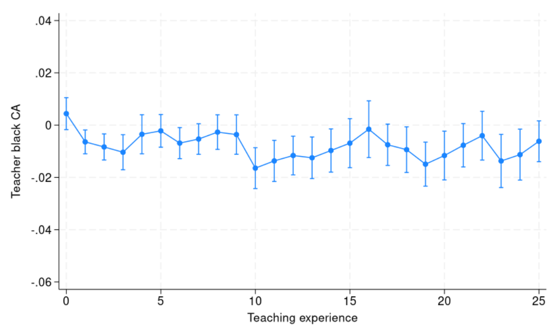
Notes: Figures show scatter plots of teacher black comparative advantage and mean value-added by subject and school level. Sample is restricted to teachers who are ever observed teaching students from both races. Mean VA is the average of race-specific VAs. Each dot is a teacher-year-subject observation. Panel A shows elementary school math, Panel B elementary school English, Panel C middle school math, and Panel D middle school English. The correlation between CA and mean VA is reported in each panel.

Figure A.6: Teaching Effectiveness-Experience Profiles

(a) Mean VA



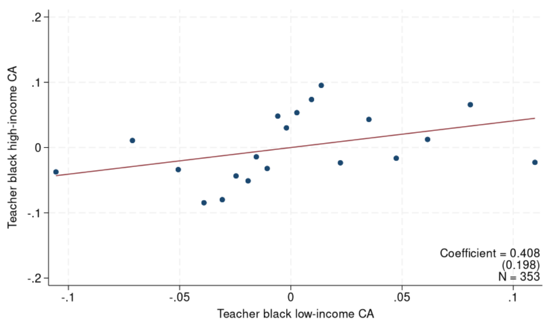
(b) Black CA



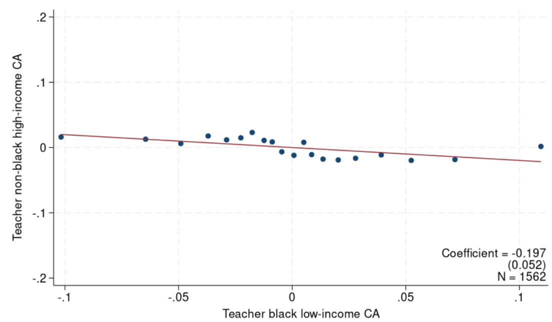
Notes: Figures show relationship between teacher black comparative advantage and mean value-added with teaching experience. Panels A and B come from regressing teacher black CA and mean VA, respectively, on fully saturated teaching experience dummies, weighted by the number of students in the classroom and clustering standard errors at the teacher level. Mean VA is the average of black- and non-black-specific VAs. Sample is restricted to teachers who are ever observed teaching students from both races.

Figure A.7: Race-by-income Teacher Comparative Advantage

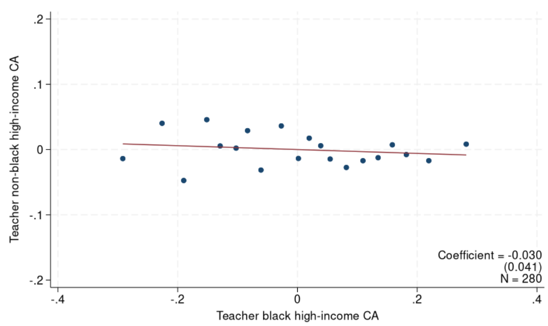
(a) Black low-income CA and black high-income CA



(b) Black low-income CA and non-black high-income CA



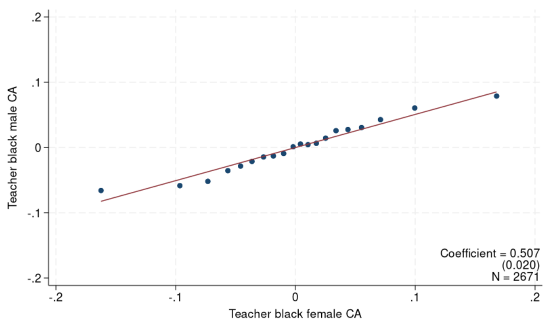
(c) Black high-income CA and non-black high-income CA



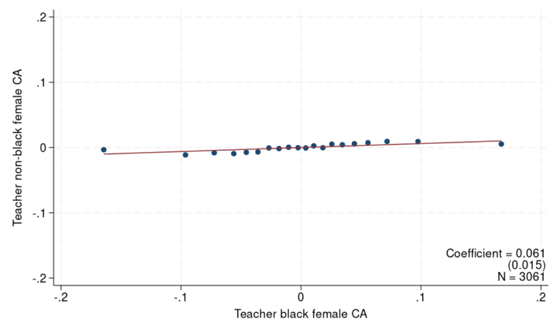
Notes: Figures show the relationship between race-by-income-specific teacher CA measures. Each figure is a binned scatterplot, where the sample is grouped into 20 equal-size groups based on values of the x-axis variable and the mean of the y-axis variable is plotted for each group. The red line is the best linear fit based on the underlying data. The reference group is non-black low-income students and each comparative advantage measure is defined in relation to this group. For instance, teacher black low-income CA is the difference between black-poor-specific teacher VA minus non-black-poor-specific teacher VA, and similarly for the other CA variables. Sample is restricted to teachers who are ever observed teaching students from the respective types used to construct each CA measure. For example, Panel A, which plots black-low-income CA and black high-income CA, restricts the sample to teachers who have taught to black low-income, black high-income and non-black low-income students. Teacher CA measures are trimmed at the top and bottom 1 percent, except for teacher black high-income CA which is trimmed at the 25 percent to deal with outliers.

Figure A.8: Race-by-gender Teacher Comparative Advantage

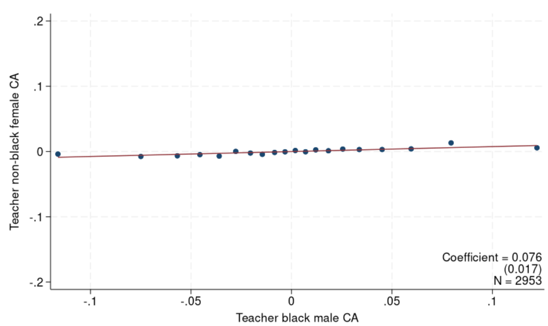
(a) Black male CA and black female CA



(b) Black male CA and non-black female CA



(c) Black female CA and non-black female CA



Notes: Figures show the relationship between race-by-gender-specific teacher CA measures. Each figure is a binned scatterplot, where the sample is grouped into 20 equal-size groups based on values of the x-axis variable and the mean of the y-axis variable is plotted for each group. The red line is the best linear fit based on the underlying data. Regressions include subject-by-school level dummies. The reference group is non-black boys and each comparative advantage measure is defined in relation to this group. For instance, teacher black female CA is the difference between black-female-specific teacher VA minus non-black-male-specific teacher VA, and similarly for the other CA variables. Sample is restricted to teachers who are ever observed teaching students from the respective types used to construct each CA measure. For example, Panel A, which plots black-female CA and black-male CA, restricts the sample to teachers who have taught to black female, black male and non-black male students. Teacher CA measures are trimmed at the top and bottom 1 percent to deal with outliers.

B Special cases of the DVA model

Previous VA models are special cases of the DVA model. I consider five cases:

(i) $K = 1$: There is a single student type. This scenario is the homogeneous VA model developed by Chetty et al. (2014a).

(ii) $\sigma_{\mu_k \mu_m, s} = 0$ for all $s \in \{1, ..., t-1\}$: Student-type-specific teacher effects are uncorrelated across student types. This implies that information from other student subgroups except for type-k students is not informative in predicting type-k-specific teacher VA. The best linear predictor for type-k-specific teacher VA becomes $\hat{\mu}_{jkt} = \boldsymbol{\psi}_k' \boldsymbol{A}_{jk}^{-t}$, and this is equivalent to separately estimating teacher effects using different subsamples.

(iii) $\mu_{jkt} = \mu_{jk}$ and $\sigma_{\mu_k \mu_m, s} = 0$ for all $s \in 1, ..., t-1$: Teacher effects are fixed and uncorrelated with each other. This scenario yields a reliability weight for type-k-specific teacher VA equal to

$$\boldsymbol{\psi}_k = \frac{\sigma_{\mu_k}^2}{\sigma_{\mu_k}^2 + \left(\sigma_{\theta_k}^2 + \sigma_{\varepsilon_k}^2 / n_k\right) / t - 1}. \tag{14}$$

This formula is discussed in Chetty et al. (2014a) as one of their special cases, and it coincides with equation 5 in Kane and Staiger (2008).

(iv) $\mu_{jkt} = \mu_{jk}$ and $K = 2$: There are two student types, and student-type-specific teacher effects are fixed but correlated with each other. The reliability weights matrix to predict type-1-specific teacher VA simplifies to a 2×1 vector of the form

$$\boldsymbol{\psi}_1 = \frac{1}{D} \begin{pmatrix} \left[\sigma_{\mu_2}^2 + \left(\sigma_{\theta_2}^2 + \sigma_{\varepsilon_2}^2 / n_2\right) / (t-1)\right] \sigma_{\mu_1}^2 - \left[\sigma_{\mu_1 \mu_2} + \sigma_{\theta_1 \theta_2} / (t-1)\right] \sigma_{\mu_1 \mu_2} \\ \left[\sigma_{\mu_1}^2 + \left(\sigma_{\theta_1}^2 + \sigma_{\varepsilon_1}^2 / n_1\right) / (t-1)\right] \sigma_{\mu_1 \mu_2} - \left[\sigma_{\mu_1 \mu_2} + \sigma_{\theta_1 \theta_2} / (t-1)\right] \sigma_{\mu_1}^2 \end{pmatrix},$$

where

$$D = \left[\sigma_{\mu_1}^2 + \left(\sigma_{\theta_1}^2 + \sigma_{\varepsilon_1}^2 / n_1\right) / (t-1)\right] \left[\sigma_{\mu_2}^2 + \left(\sigma_{\theta_2}^2 + \sigma_{\varepsilon_2}^2 / n_2\right) / (t-1)\right] - \left[\sigma_{\mu_1 \mu_2} + \sigma_{\theta_1 \theta_2} / (t-1)\right]^2.$$

This formula is similar to the optimal weights derived by Lefgren and Sims (2012) in their equation 7, with some differences. They consider the case when teachers affect two student outcomes (e.g., math and English), while I consider the case when teachers affect multiple student types; therefore, the assumptions around the individual- and classroom-level shocks slightly vary.

(v) $\mu_{jkt} = \mu_{jk}$ and $\sigma_{\theta_m\theta_n} = 0$: Student-type-specific teacher effects are fixed but correlated with each other, and student-type-specific classroom shocks are uncorrelated with each other. Under this scenario, the reliability weights matrix can be expressed as

$$\boldsymbol{\psi} = (\boldsymbol{T} + \boldsymbol{V})^{-1}\boldsymbol{T}, \tag{15}$$

where

$$\boldsymbol{T} = \begin{pmatrix} \sigma^2_{\mu_1} & \cdots & \sigma_{\mu_1\mu_K} \\ \vdots & \ddots & \vdots \\ \sigma_{\mu_K\mu_1} & \cdots & \sigma^2_{\mu_K} \end{pmatrix}$$

and $\boldsymbol{V}$ is a diagonal matrix with its k-th element equal to $\left(\sigma^2_{\theta_k} + \frac{\sigma^2_{\varepsilon_k}}{n_k}\right)/t{-}1$. Equation 15 coincides with equation 3.57 of Raudenbush and Bryk (2002)'s hierarchical model.

C Sources of Heterogeneity in Teacher CA

This section examines what teacher characteristics and teaching practices are associated with greater CAs for black students.

C.1 Teacher CA and teacher characteristics

Given the existing racial achievement gaps, teachers with greater black CA would contribute more to racial equity. Appendix Figure A.5 shows a scatterplot of teachers' comparative and absolute advantage measures by subject and school level, where absolute advantage is the average of their black- and non-black-specific teacher VA. I use mean VA rather than the reference-group-specific VA for this exercise to avoid the mechanical negative correlation between teacher CA and reference-group VA. Both mean teacher VA and reference-group VA have a one-to-one relationship with homogeneous teacher VA (see Appendix Table A.5). Teachers with similar absolute advantage have different degrees of CA for black students, and, interestingly, the relationship between CA and mean VA varies markedly across subjects and school levels. Relative to the average teacher, high-equity teachers are those in quadrants I and II, and high-effective teachers are those in quadrants I and IV.

[Appendix Figure A.5 about here]

To identify high-equity teachers, I estimate the following teacher-level regression:

$$\mathbb{I}[CA_{jt} > 0] = X'_{jt}\beta + \varepsilon_{jt}, \tag{16}$$

where $\mathbb{I}[CA_{jt} > 0]$ indicates whether teacher j's black CA is greater than the average teacher's, X_{jt} is a vector of teacher characteristics, and ε_{jt} is the error term. Given that a teacher may increase her black CA by *reducing* her VA for non-black students rather than *increasing* her black-specific VA, I explore if the same teacher characteristics are correlated with being highly effective. I identify high-effective teachers and high-effective, high-equity teachers by replacing the dependent variable for $\mathbb{I}[meanVA_{jt} > 0]$ and $\mathbb{I}[meanVA_{jt} > 0, CA_{jt} > 0]$, respectively. Appendix Table A.8 presents the results, clustering standard errors at the teacher level and restricting the sample to teachers who are ever observed teaching students from both racial groups.

[Appendix Table A.8 about here]

I find that few teacher characteristics are strongly associated with the likelihood of being a high-equity teacher. Being young and having more years of experience and a master's degree are positively related with the likelihood of being highly effective (Column 1 of Appendix Table A.8). However, having a master's degree and more years of teaching experience decreases the likelihood of being a high-equity teacher (Column 2), which indicates that teachers with experience and graduate degrees increase test scores but much less for black students. Overall, teacher characteristics explain little of the variation in teacher CA.

Additionally, I find that if a teacher is black, they are 3.4 percentage points more likely to be a high-equity teacher, but it is not statistically significant (Column 2). They are also 5 percentage points more likely to be high-effective, high-equity teachers (Column 3). The insignificant relationship with teacher CA seems to contradict the well-established finding of positive racial match effects when students are assigned to same-race teachers. Studies on this topic use different research designs to estimate racial match effects. I reproduce their findings using the design best suitable for my data by following Egalite et al. (2015) and estimating the student fixed effect model.[33] I find a statistically significant racial match effect. However, when I disaggregate the racial match indicator into black race match and non-black race match indicators, I find that the result is driven by the latter indicator and the black race match coefficient is indistinguishable from zero at conventional levels (see Appendix Table A.11). Therefore, the finding that black teachers do not have significantly higher CA for black students may reflect the result in this setting that the black race match coefficient is indistinguishable from zero.[34]

[33]Studies on student-teacher demographic match effects use different research designs to estimate racial match effects; for example, Dee (2004) exploits the random assignment of students to teachers, Egalite et al. (2015) employ a student fixed effect design, Delhommer (2022) uses a course fixed effects design given that courses have multiple sections with different teachers, and Gershenson et al. (2022) employ a specification with the share of black teachers. I estimate the following fixed effect model:

$$A_{it} = \beta_0 + X'_{jt}\beta_1 + \gamma RaceMatch_{ijt} + \phi GenderMatch_{ijt} + \alpha_i + \varepsilon_{it}, \tag{17}$$

where A_{it} is student i's residualized test scores, X'_{jt} includes teacher characteristics, $RaceMatch_{ij}$ indicates whether the student's race (black or non-black) matches that of her teacher, $GenderMatch_{ij}$ indicates whether the student's gender (female or male) matches that of her teacher, α_i is a student fixed effect, and ε_{it} is the error term clustered at the cohort level. The results are presented in Appendix Table A.11. This specification exploits within-student variation when the same student is assigned to teachers with a similar or different race (and gender).

[34]Another explanation of the seemingly contradiction is that grouping Hispanic students into the non-black group category may mute the racial match effects because both black and Hispanic students benefit from having minority teachers. To explore this, I re-estimate race-specific teacher VA for three racial groups (black, Hispanic, and white plus

Next, I investigate the experience profile of teacher quality in Appendix Figure A.6. Panel A shows the experience profile for mean teacher VA and Panel B for teacher black CA by regressing each teacher quality measure on fully saturated experience dummies with 25 years of experience or more lumped in one category. Mean teacher VA has a positive trend with experience, while black CA declines. This suggests that the experience itself does not translate into higher effectiveness for closing racial achievement gaps. An alternative explanation is negative selection where teachers with higher black CA are more likely to exit the profession, as in Wiswall (2013).

[Appendix Figure A.6 about here]

C.2 Teacher CA and teacher behaviors

To examine what teaching practices are associated with greater CAs for black students, I link evaluators' and students' ratings of various teaching practices with the teachers' CA measure. Starting with evaluators' classroom observation ratings, I estimate the following teacher-level model:

$$O_{jt} = \lambda CA_{jt} + \phi VA_{jt} + X'_{jt}\gamma + \varepsilon_{jt}, \tag{18}$$

where O_{jt} is the evaluator-assigned score for teacher j in year t; CA_{jt} and VA_{jt} are one-year-leave-out forecasts of teacher black CA and reference-group VA; X_{jt} is a vector of teacher characteristics, which I include in some specifications to control for observable sorting of teachers to evaluators; and ε_{jt} is the error term. I rescale teacher CA and VA using their estimated variance (i.e., CA_j/σ_{CA} and VA_j/σ_{VA}) to ease interpretation, so that a 1-unit increase means a 1-standard deviation increase. This analysis restricts the sample to teachers who are ever observed teaching both black and non-black students so that teacher CA is estimated with the data instead of using counterfactual estimates.

Appendix Table A.9 shows the results where odd-numbered columns exclude teacher characteristics and even-numbered columns include them. Standard errors are clustered at the teacher level and are reported in parentheses. The dependent variables are the average classroom observation

other races) and compute teacher CA for black students relative to white students. Similar to my previous finding, I find a null association between teacher race and teacher CA for black students when the comparison group is white students, although the estimate is noisier due to smaller sample sizes when restricting the sample to classrooms with both black and white students.

score in each of the four domains—planning and preparation, classroom environment, instruction, and professional responsibilities—and each average score is standardized to have mean zero and standard deviation one by subject-school level-year. The table shows that a 1σ increase in teacher reference-group VA is associated with 0.2–0.29σ higher scores across all four teaching practices, and these associations are statistically significant. However, teacher black CA is not strongly associated with any of the evaluated teaching practices.

[Appendix Table A.9 about here]

Next, I use student ratings of teaching practices to test whether students identify teachers with greater CA for black students. The dependent variable is now the student-level average of survey ratings, $O_{jt} = \frac{1}{N_{jt}} \sum_{i:j=j(c(i,t))} O_{it}$, where O_{it} is student i's rating of her teacher and N_{jt} is the number of students in the teacher's classroom who completed the survey. To construct the dependent variable, I first group students' item responses into seven indexes—peer support, classroom rigor, academic press, course clarity, academic engagement, academic personalism, and classroom disruptions (see Appendix Table A.2)—by taking the student-level average of the items within these groups. I then take the teacher-level average of the student survey indexes. Last, I standardize each teacher-level average score to have mean 0 and standard deviation 1 by subject-school level-year.

Appendix Table A.10 shows the regression results, where the dependent variable is each of the student survey indexes and observations are weighted by the number of students who completed the survey in each classroom. Standard errors are clustered at the teacher level and are reported in parentheses. Panel A shows results for specifications without teacher controls and Panel B includes these controls. The results indicate that a 1σ increase in teacher reference-group VA is significantly associated with 0.08–0.17σ higher student ratings across all indexes (Columns 1–7). Teachers with higher black CA receive tend to receive higher ratings across domains but they are statistically indistinguishable from zero. In sum, although not statistically significant, there is suggestive evidence that teacher CA may be malleable by teaching practices that were rated by students.

[Appendix Table A.10 about here]

C.3 Possible drivers of variation in teacher CA for black students: differential effectiveness in income and other characteristics.

Black students tend to come from economically disadvantaged families, and therefore the observed differential teacher impacts by race may be driven by differential teacher effectiveness by income level or other characteristics. To address this concern, I estimate race-by-income-specific teacher VA and relate these estimates between them. This interaction creates four student subgroups: low-income black, high-income black, low-income non-black, and high-income non-black.

Since the majority of students in Chicago Public Schools are eligible for free/reduced-price lunch status, I set the low-income non-black group as the reference group and construct teacher CA for the other student subgroups with respect to the reference group. I obtain three measures of teacher CA: low-income black, high-income black, and high-income non-black. If racial disparities are driven by income disparities, then one should observe a high correlation between high-income black CA and high-income non-black CA but observe a low correlation between low-income black CA and high-income black CA.

Appendix Figure A.7 plots these correlations for elementary school teachers, the subsample of teachers for which socioeconomic-status-specific teacher CA passes the quasi-experimental test of forecast bias (see Section 7.4). Panel A of Appendix Figure A.7 correlates estimates of teacher low-income black CA and high-income black CA for the sample of teachers who are ever observed teaching low- and high-income black students and low-income non-black students since these groups were used to construct these measures. The figure is a binned scatterplot where observations are grouped into 20 equally sized groups based on values of the x-axis variable, and their average of the y-axis variable is plotted. Bivariate regression results are reported with standard errors clustered at the teacher level. Panels B and C present similar bivariate regressions for the pairs specified in each graph. The results show that black-specific CA measures are positively correlated between them (Panel A) but not with income-specific CA (Panels B and C).

[Appendix Figure A.7 about here]

I next investigate whether this pattern occurs with other student demographics. I repeat this exercise by estimating race-by-gender-specific teacher effects, setting non-black boys as the reference group and constructing black male CA, black female CA, and non-black female CA relative to this reference group. The results are similar in that the black-specific CA measures are correlated between them but not with gender-specific CA (see Appendix Figure A.8).

D Welfare consequences of considering teacher CA

In this section, I conduct three sets of counterfactual policy simulations to quantify the potential efficiency and equity gains from incorporating teacher CA into policy decisions. The simulated policies reflect different ways in which information on teacher black CA could be used. For simplicity, I focus on elementary school math teachers and their impacts on racial achievement gaps in math test scores. For consistency across simulations, I restrict this sample to classrooms with similar class sizes (those within the 25th and 75the percentiles of the teacher-subject-year cell size distribution) as the optimal teacher-to-classroom allocation policy discussed below assumes equal class sizes.

Contrary to the homogeneity assumption in teacher effects, the classroom composition matters under disparate teacher effects. A teacher's total output produced in a classroom (i.e., her total increase in test scores) is a linear combination of her race-specific effects weighted by her classroom's racial composition. More formally, let m_{j0t} and m_{j1t} be teacher j's true VA for non-black and black students in year t, respectively. Let p_c be the proportion of black students in classroom c. The classroom output in per student terms produced by teacher j when she is assigned to classroom c in year t, $Q_{jt}(p_c)$, is

$$Q_{jt}(p_c) = (1 - p_c) \times m_{j0t} + p_c \times m_{j1t} = VA_{jt} + p_{jc} \times CA_{jt}, \tag{19}$$

where $VA_{jt} = m_{j0t}$ is the reference-group teacher VA and $CA_{jt} = m_{j1t} - m_{j0t}$ is teacher black CA. In what follows, the expected impact of counterfactual policies makes use of this expression.

D.1 Teacher accountability policy

D.1.1 Benchmark policy: homogeneity-based teacher performance measure.

A policy often discussed in the literature is to replace teachers in the bottom 5 percent of the homogeneous teacher VA distribution with teachers of average quality (Hanushek, 2009). Given that classroom racial composition matters due to race-specific teacher impacts, this homogeneity-based policy may fire teachers who would otherwise perform better in a typical classroom while keeping others who would otherwise perform worse. For example, a teacher with a large negative VA for black students may stay because she happens to be in a classroom with a low proportion of black students. To level the playing field, the counterfactual accountability policy deselects teachers based on their impacts in a typical classroom, holding the racial composition the same for all teachers.

To replicate the homogeneity-based benchmark policy when teacher effects are heterogeneous, I assume that a teacher's estimated homogeneous VA is her classroom output given the observed classroom assignment, $Q_{jt}(p_c^0)$, where p_c^0 is the proportion of black students in her observed classroom. This section builds on Chetty et al. (2014b)'s online Appendix D.

Selection on true race-specific teacher VA. This scenario occurs when the true race-specific teacher VAs are known, that is, when we can predict with certainty future teacher impacts. The expected gain per student from replacing teachers in the bottom 5 percent of the accountability measure $mjjt$ with teachers of average quality is

$$G(0) = -\mathbb{E}\left[Q_{jt}(p_c) \mid Q_{jt}(p_c) < F_{Q_p}^{-1}(0.05)\right], \tag{20}$$

where the expected value of Q_{jt} depends on classroom assignment and is conditional on the teacher's accountability measure (Q_{jt}) being below the 5th percentile. I calculate the gains from deselection based on true race-specific VA using analogous Monte Carlo simulations that I explain in more detail in the next subsection, except that I draw scores from the VCV matrix of true VA, $\mathbf{\Sigma}_\mu = \left(\begin{smallmatrix} \Sigma_{\mu_0} & \Sigma_{\mu_0\mu_1} \\ \Sigma_{\mu_0\mu_1} & \Sigma_{\mu_1} \end{smallmatrix}\right)$ instead of test scores, $\mathbf{\Sigma}_A$. The off-diagonal elements of their submatrices are identical, but the diagonal elements of Σ_{μ_0} and Σ_{μ_1} reflect only the variance of teacher quality $\sigma^2_{\mu_0}$ and $\sigma^2_{\mu_1}$, respectively, and the diagonal elements of $\Sigma_{\mu_0\mu_1}$ reflect the within-year covariance $\sigma_{\mu_0\mu_1}$. I use quadratic estimates of these parameters reported in the last rows of Table 3 for this simulation.[35]

Appendix Table A.12 presents estimates of the efficiency and equity gains of this policy (row 1), where the gains are expressed in test score standard deviation units per student. The classroom output increase per student for the 5 percent of classrooms affected by this policy is in Column 1. Column 2 takes in to account all affected and unaffected classrooms, and Columns 5 and 6 decompose these gains into average test score gains for non-black students (ΔA_0) and black students

[35] The VCV matrix needs to be a Hermitian, positive-definite matrix in order to apply the Cholesky decomposition for the simulations. Therefore, I use the lower bound estimate of the within-year covariance (0.040) rather than the quadratic estimate (0.048) so that the VCV matrix has a Cholesky decomposition.

(ΔA_1). Column 3 expresses the efficiency gains as a percentage of the standard deviation of teacher absolute advantage (average of race-specific teacher VA) to compare it with a hypothetical policy that increases average teacher quality. Column 4 shows the efficiency gains as a percentage of the benchmark policy's impact. Column 6 shows the equity impacts (the difference between Columns 5 and 6), where a positive value reflects an increase in the racial achievement gap and a negative value reflects a decrease. Column 8 expresses these equity gains as a percentage of the black-non-black achievement gap.

[Appendix Table A.12 about here]

Replacing the lowest-performing teachers based on their true VAs with average teachers would increase test scores by 0.508σ per affected student on average (row 1, Column 1).

Selection on estimated race-specific teacher VA. A more realistic scenario is deselecting teachers based on estimated impacts since the true effects are not observed. I add the following notation to help with exposition. Let $\hat{Q}_{j,n+1}(p_c; n)$ be teacher j's predicted impact in year $n+1$ based on test score data from years $t = 0, \ldots, n$ and given her future classroom assignment (whose racial classroom composition is p_c). This term is given by

$$\hat{Q}_{j,n+1}(p_c; n) = (1 - p_c) \times \hat{m}_{j0,n+1} + p_c \times \hat{m}_{j1,n+1} = \widehat{VA}_{j,n+1} + p_c \times \widehat{CA}_{j,n+1}, \tag{21}$$

where $\hat{m}_{j0,n+1}$ and $\hat{m}_{j1,n+1}$ are out-of-sample predictions of non-black- and black-specific teacher VA, and $\widehat{VA}_{j,n+1}$ and $\widehat{CA}_{j,n+1}$ are constructed based on these predictions.

The gain in year $n+1$ from replacing the bottom 5 percent of teachers with average teachers, based on their homogeneous accountability measure based on test score data from years $t = 1, \ldots, n$, is

$$G(p^0; n) = -\mathbb{E}\left[Q_{j,n+1}(p_c) \mid \hat{Q}_{j,n+1}(p_c^0; n) < F^{-1}_{\hat{Q}(p^0;n)}(0.05)\right], \tag{22}$$

where the expected value of the true impacts $Q_{j,n+1}$ is conditional on the teacher's predicted homogeneous VA falling below the 5th percentile, and predictions are based on n years of data. I estimate the impacts for $n = 3$ (three years of data) given the finding that the marginal gains from obtaining one more year of data are outweighed by the expected cost of keeping a low-VA teacher for one additional year (Chetty et al., 2014b; Staiger and Rockoff, 2010). This policy has been estimated to increase students' lifetime earnings by \$250,000 per treated classroom (Chetty et al., 2014b).

The Monte Carlo simulations in this setting works as follows. I simulate teachers' disparate VA under the assumption that μ_{j0t} and μ_{j1t} follow a multivariate normal distribution. First, I construct $\mathbf{\Sigma}_A$, the VCV matrix of $\boldsymbol{A}_j^{-t} = (\boldsymbol{A}_{j0}^{\prime -t} \boldsymbol{A}_{j1}^{\prime -t})'$, the vector of past black- and non-black-specific class average scores, using the parameters of the autocovariance and crosscovariance vectors of test scores reported in Columns 1–3 of Table 3. The VCV matrix $\mathbf{\Sigma}_A = \left(\begin{smallmatrix} \Sigma_{A_{00}} & \Sigma_{A_{01}} \\ \Sigma_{A_{10}} & \Sigma_{A_{11}} \end{smallmatrix}\right)$ is a block matrix whose off-diagonal submatrices $\Sigma_{A_{01}} = \Sigma_{A_{10}}$ are identical. I define the off-diagonal elements of $\Sigma_{A_{00}}$ and $\Sigma_{A_{11}}$ based on the race-specific autocovariances reported in Table 3, setting the autocovariance $\sigma_{Aks} = \sigma_{Ak4}$ for $s > 4$. Similarly, I define the off-diagonal elements of $\Sigma_{A_{01}}$ based on the crosscovariances reported in Table 3. I define the diagonal elements of $\Sigma_{A_{00}}$ and $\Sigma_{A_{11}}$ as the variance of race-specific mean class test scores, which I compute based on the estimates in Table 3 as $\sigma^2_{\mu_k} + \sigma^2_{\theta_k} + \sigma^2_{\varepsilon_k}/N_k$, where $N_0 = 25$ and $N_1 = 15$ are obtained assuming an average number of students per class of $N_c = 38.9$ and proportion of black students of $p_c = 0.37$. Similarly, I define the diagonal elements of $\Sigma_{A_{01}}$ as the covariance of mean class test scores, which I compute as $\sigma_{\mu_k \mu_m} + \sigma_{\theta_k \theta_m}$ reported in Table 3.

Second, I simulate draws of average race-specific class scores from a $N(\mathbf{0}, \mathbf{\Sigma}_A)$ distribution for one million teachers and calculate $\hat{m}_{j0,n+1}$ and $\hat{m}_{j1,n+1}$ based on test scores from the first n periods using the same method used to construct race-specific VA estimates. Additionally, I take the observed classroom racial composition as given, create one million classrooms and randomly assign them to teachers. Finally, I calculate the conditional expectation in Eq. 22 as the mean classroom output per student in year $n+1$, $Q_{j,n+1}(p_c)$, for teachers with accountability measure $m_{j,n+1}$ (explained below) in the bottom 5 percent of the distribution and given classroom assignment p_c.

Appendix Table A.12 presents estimates of the expected efficiency and equity gains of this policy (row 2). Replacing the lowest-performing teachers based on their estimated homogeneous VA with average teachers would increase test scores by 0.365σ per student on average (row 2, Column 1 of Table A.12). This is 30 percent lower than the scenario when the true teacher VA is observed (row 1, Column 1). Since students in unaffected classrooms see no changes in their test scores, the average test score impact of this policy across affected and unaffected classrooms is $0.0.18\sigma$ per student, where non-black students see an average test score gain of 0.019σ and black students of 0.016σ.

The efficiency gain is equivalent to an overall improvement in teacher absolute advantage quality of 4 percent. However, this policy would increase the racial achievement gap by only 0.004σ, which is equal to 0.4 percent of the black-white achievement gap and 0.9 percent of the black-non-black gap (Columns 7 and 8).

D.1.2 Heterogeneity-based teacher performance measure

The heterogeneity-based performance measure aims to level the playing field by mesauring teachers' performance in a typical classroom.

Selection on estimated race-specific teacher VA. I simulate a policy that incorporates teachers' disparate impacts by employing the accountability measure $\hat{Q}_{jt}(p; n)$, where p is the expected proportion of black students in a randomly chosen classroom.[36] Here, all teachers are evaluated as if they were teaching in a representative classroom, hence leveling the playing field. When teachers are deselected based on their estimated teacher VA, the expected efficiency gain is $G(p; n)$ as in Eq. 22, where the estimated impacts are based on n years of data and teachers are evaluated as if teaching in a representative classroom.

This policy would increase total output by an average of 0.018σ per student (row 3, Column 3 of Appendix Table A.12) and increase the racial achievement gap by only 0.004σ or 0.9 percent of the black-non-black achievement gap (row 3, Columns 7 and 8). Relative to the benchmark policy, the heterogeneity-based accountability policy would increase efficiency by an additional 1.2 percent, which in monetary terms would be equivalent to \$3,100 ($= \$250,000 \times 1.2\%$) per classroom.

D.2 Teacher-to-classroom reassignment policy

A social planner (e.g., school principal) could use information on teacher CA to better match teachers to students to maximize efficiency and equity. Under the homogeneity of the teacher effects assumption, reallocating teachers across classrooms is a zero-sum game as the increase in student test scores when a low-VA teacher is replaced by a high-VA one is met by an equal-sized decrease in test scores when the high-VA teacher leaves her classroom. On the other hand, under the heterogeneity assumption of teacher effects, teachers have varying CAs as different teachers are more

[36] Other teacher performance measures may include, (i) mean VA, $m_{j,n+1} = (\hat{\mu}_{j0,n+1} + \hat{\mu}_{j1,n+1})/2$, and (ii) the expected classroom impact $m_{j,n+1} = \hat{Q}_{j,n+1}(p_c^*)$ when the teacher's classroom assignment is expected to be p_c^*.

effective at raising the test scores of different student subgroups. Therefore, reallocating teachers across classrooms becomes a positive-sum game, and better matching teachers to classrooms could increase student test scores overall.

To formally see this, suppose teacher j is switched with teacher k, while their classroom composition stays the same. The impact of this switch on total output per student is

$$\underbrace{Q_{jt}(p_k) - Q_{kt}(p_k)}_{\Delta\text{output teacher } j \text{ replaces teacher } k} \quad + \quad \underbrace{Q_{kt}(p_j) - Q_{jt}(p_j)}_{\Delta\text{output teacher } k \text{ replaces teacher } j} \quad = (p_k - p_j)(CA_{jt} - CA_{kt}).$$

There are efficiency gains of reallocation if $p_k > p_j$ and $CA_{jt} > CA_{kt}$; that is, the teacher with the highest black CA is assigned to the classroom with the largest proportion of black students. The reallocation policy is in theory resource neutral since it does not require hiring additional teachers; however, in practice it may require resources.

D.2.1 Reallocating teachers to maximize student achievement

In this exercise, the social planer's goal is to maximize efficiency (i.e., maximize student achievement or total output) by reallocating teachers to classrooms. Her utilitarian social welfare function in a given year (where subscript t is omitted for simplification) is

$$W = \sum_i \mathbb{E}[A_i], \tag{23}$$

where A_i is student test scores and the expected value is conditional on teacher assignment. Classroom racial composition is taken as given.[37] For simplicity I omit subscript t, but the maximization problem is solved each year.

Substituting for the components that form test score residuals, $A_i = \mu_{jk(i)} + \theta_{ck(i)} + \varepsilon_i$ and assuming the social planner does not know the realizations of classroom and individual shocks at the moment of her decision, the social welfare function simplifies to

$$W = \sum_i \mu_{jk(i)}$$

[37]Another potential policy is to reallocate students to classrooms rather than teachers to classrooms, but this entails accounting for and quantifying peer effects. This is out of scope of this paper.

.

Let N_{ck} denote the number of students of type k in classroom c. Given that students are nested within classrooms and each classroom has a teacher, W can be reexpressed as

$$W = \sum_c \sum_k N_{ck}\mu_{jk}$$

where the sums are over classrooms and student types. Assume $k \in \{0, 1\}$ and let $N_c = N_{c0} + N_{c1}$ be class size. Assume also equal class size $N_c = N$ but the composition of classrooms is allowed to vary. The welfare function can be expressed in per pupil terms as

$$W^* = \sum_c [p_c CA_j] + \sum_j \mu_{j0}$$

where $W^* = W/N_T$, $N_T = CN$ is total number of students, $CA_j = \mu_{j1} - \mu_{j0}$ is teacher j's comparative advantage, and $p_c = N_{c1}/N$ is the proportion of type-1 students in classroom c.

Proposition 1 (Optimal teacher allocation) *Given classrooms' student composition, the optimal assignment policy ranks teachers based on their CA and classrooms based on their proportion of type k students, and match them accordingly, resulting in positive assortative matching.*

Choose two teachers A and B, with $CA_A > CA_B$, and two classrooms X and Y, with $p_X > p_Y$. Then, the optimal allocation assigns teacher A to classroom X and teacher B to classroom Y. Note that the assumption about difference in CA is the single crossing condition:

$$\mu_{A1} - \mu_{B1} > \mu_{A0} - \mu_{B0}$$

Proof: Assume otherwise the optimal allocation assigns teacher A to classroom Y and teacher B to X. This produces total out per student equal to:

$$\underbrace{p_Y\mu_{A1} + (1 - p_Y)\mu_{A0}}_{\text{Teacher A's output in class Y}} + \underbrace{p_X\mu_{B1} + (1 - p_X)\mu_{B0}}_{\text{Teacher B's output in class X}}$$

$$= p_Y(\mu_{A1} - \mu_{A0}) + \mu_{A0} + p_X(\mu_{B1} - \mu_{B0}) + \mu_{B0}$$

$$> p_X(\mu_{A1} - \mu_{A0}) + \mu_{A0} + p_Y(\mu_{B1} - \mu_{B0}) + \mu_{B0}$$

The second line rearranges terms, and third line is the definition of optimal assignment (output greater than other assignments). Rearranging terms I get:

$$
\begin{aligned}
(p_X - p_Y)(\mu_{B1} - \mu_{B0}) > (p_X - p_Y)(\mu_{A1} - \mu_{A0}) &\Rightarrow \\
(\mu_{B1} - \mu_{B0}) > (\mu_{A1} - \mu_{A0}) &\Rightarrow \\
CA_B > CA_A &
\end{aligned}
$$

which is a contradiction of assumption $CA_A > CA_B$. $\square$

The optimal teacher-to-classroom reallocation policy matches teachers with greater black CA to classrooms with greater proportions of black students, generating positive assortative matching. I reallocate teachers within a subject-grade level-school year; e.g., I switch 4th grade math teachers with other 4th grade math teachers in year 2017. As suggested by the quasi-experimental evidence, this policy assumes that estimated teacher CA predicts teachers' disparate effects when teachers are reallocated across classrooms within schools or across schools, even for moves not observed in the data.

Reallocation based on estimated teacher CA. A more realistic scenario is to reallocate teachers based on estimated impacts since the true effects are not observed. Under this scenario, the social planner matches teachers to classrooms based on $\widehat{CA}_{j,n+1} = \hat{m}_{j1,n+1} - \hat{m}_{j0,n+1}$, where $\hat{m}_{jk,n+1}$ comes from noisy estimates of non-black- and black-specific teacher VA as described in the teacher accountability policy section (Appendix Section D.1).

Row 5 of Table A.12 shows the efficiency and equity gains from reallocating teachers within schools to maximize total output, and row 6 shows the results when teachers are reallocated across schools. Within-school teacher reallocation produces small efficiency gains (less than 0.001σ, or 0.6 percent of the benchmark policy). However, across-school reallocation produces larger efficiency gains of 0.003σ per student on average, which represents about 16 percent of the benchmark policy's impact. In monetary terms, the gain is equal to \$40,500 ($= \$250,000 \times 16.2\%$) per classroom. Reallocating teachers to maximize student achievement would come at the cost of widening the racial achievement gap by 0.038σ, which is 9.2 percent of the black-non-black gap.

D.2.2 Reallocating teachers to minimize racial achievement gaps

The social planner may care about reducing racial achievement gaps. To quantify the maximum equity gains from reallocating teachers, the social planner's problem is to minimize the following welfare function:

$$W = \frac{1}{N_0} \sum_{i:k(i)=0} \mathbb{E}[A_i] - \frac{1}{N_1} \sum_{i:k(i)=1} \mathbb{E}[A_i], \tag{24}$$

where N_0 and N_1 are the total number of non-black and black students, respectively, and A_i is student test scores and the expected value is conditional on teacher assignment. The first term is the average expected test scores of non-black students, and the second term is the average for black students; hence the difference is the black-non-black racial achievement gap. This minimization problem has no simple closed-form solution, but one way to achieve this goal is to assign teachers with greater absolute advantage, i.e., $meanVA = (\mu_{j1} + \mu_{j0})/2$, to classrooms with larger proportions of black students (p_c).

Reallocation based on estimated teacher absolute advantage. Here again, a more realistic scenario is to reallocate teachers based on estimated impacts since the true effects are not observed. Under this scenario, the social planner matches teachers to classrooms based on $mean\widehat{VA}_{j,n+1} = (\hat{\mu}_{j1,n+1} + \hat{\mu}_{j0,n+1})/2$, where $\hat{\mu}_{jk,n+1}$ comes from noisy estimates of non-black- and black-specific teacher VA.

Row 6 of Appendix Table A.12 shows results when teachers are reallocated across schools. It indicates a maximum reduction in the racial achievement gap by 0.150σ, which is 36.2 percent of the black-non-black achievement gap. This equity gain comes with a trade-off of reducing efficiency by -0.005σ because the increase in black students' scores does not fully compensate the reduction in non-black students' scores.

Next, I perform a similar gap-minimizing reallocation exercise subject to not decreasing total students' test scores. That is, the optimization problem is subject to $\frac{1}{N_0} \sum_{i:k(i)=0} \mathbb{E}[A_i] + \frac{1}{N_1} \sum_{i:k(i)=1} \mathbb{E}[A_i] \geq A^0$, where A^0 is the observed average test score. This constrained minimization problem has no simple closed-form solution, but I use the following algorithm to achieve this goal. First, calculate the observed average student test scores and set it as the target value. Second, sort teachers based on their absolute advantage and classrooms based on their proportion of black students and match them accordingly (as in the unconstrained teacher reallocation simulation exercise).

Then, I calculate again students' average test scores under the counterfactual allocation. If the counterfactual test score average is greater than or equal to the observed target value, stop; otherwise, continue to the next step. Fourth, to meet the target value, re-sort teachers based on their black CA, and allocate one-Xth of them to classrooms with the highest proportions of black students, where X is the number of quantiles that I set to 1,000. As discussed above, allocating teachers based on their CA maximizes the efficiency to meet the target value. Fifth, these one-Xth of teachers form the "do not reallocate" group. Repeat steps 2 to 5, leaving the do not reallocate group intact and reallocating the remaining teachers until the target value is met.

The last row of Appendix Table A.12 indicates that the constrained reallocation policy would reduce the black-non-black gap by 3.6 percent, without reducing average test scores. These results point to the potential of increasing the efficiency and equity of education systems when teacher CA is considered in policy decisions.

D.3 Professional development to reduce racial disparities

Information on teacher CA could be used to target resources for teachers who need them the most. For instance, one could provide professional development to teachers who are widening racial achievement gaps, akin to teacher training aimed at reducing implicit and explicit racial bias or at addressing other sources that create racial disparities. The counterfactual policy selects teachers in the bottom 5 percent of the teacher black CA distribution, akin to the teacher accountability policy discussed above, and replaces their CA quality to that of the average teacher (which is zero) holding their reference-group VA fixed. This policy does not entail replacing teachers but reducing their disparate impacts. It assumes that teacher CA is malleable, at least for teachers in the bottom 5 percent of the black CA distribution.

Ceteris paribus, changing a teacher's CA would only impact black students. The change in output per student produced by teacher j in classroom c when her CA changes to zero can be expressed as

$$R_{jt}(p_c) = Q_{jt}(0) - Q_{jt}(p_c) = -p_c \times CA_{jt}. \tag{25}$$

This policy simulation follows closely the teacher accountability simulation with the exception that teacher j's impact of changing her black CA to that of the average teacher in year t is $R_{jt}(p_c)$. Here again, we need to distinguish between selection based on true and estimated teacher CA.

Selection on estimated teacher CA. In this scenario a teacher's true CA cannot be observed. Let $\widehat{CA}_{j,n+1}$ be teacher j's estimated CA for year $n+1$ based on test score data from years $t = 1, \ldots, n$. The expected gain of providing professional development to teachers in the bottom 5 percent of the CA distribution is

$$H(p_c; n) = -\mathbb{E}\left[p_c \times CA_{j,n+1} \mid \widehat{CA}_{j,n+1} < F^{-1}_{\widehat{CA}}(0.05)\right]. \tag{26}$$

I calculate this expected value using Monte Carlo simulations, similar to that described in Appendix Section D.1 for the teacher accountability policy. I calculate the gains of this policy based on three years of data for making forecasts.

Reducing racial disparities through professional development offered to the bottom 5 percent of the black CA quality distribution would increase average test scores by 0.038σ per student in affected classrooms. The efficiency gains would be equal to 10.4 percent of the benchmark policy or \$25,950 ($= \$250,000 \times 10.4\%$) per affected classroom, and the racial achievement gaps would be reduced by 0.7–1.5 percent.