NBER WORKING PAPER SERIES

RACE AND THE MISMEASURE OF SCHOOL QUALITY

Joshua Angrist
Peter Hull
Parag A. Pathak
Christopher R. Walters

Working Paper 29608
http://www.nber.org/papers/w29608

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2021

Our title is a tribute to Gould (1981), who immeasurably shaped the 20th century discussion of race in social science. Financial support from the Walton Foundation and the National Science Foundation is gratefully acknowledged. We thank the staff of GreatSchools.org for many helpful conversations. Thanks also to Eryn Heying, Anna Vallee, and Talia Gerstle for invaluable administrative support. Jimmy Chin, Ray Han, Russell Legate-Yang, and Ini Umosen provided expert research assistance. The research described here was carried out under data-use agreements between MIT and the NYC and Denver Public School districts. The IRB of record was MIT COUHES, under protocol 1904808765. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Race and the Mismeasure of School Quality
Joshua Angrist, Peter Hull, Parag A. Pathak, and Christopher R. Walters
NBER Working Paper No. 29608
December 2021
JEL No. I21,I24,I26,I28

ABSTRACT

In large urban districts, schools enrolling more white students tend to have higher performance ratings. We use an instrumental variables strategy leveraging centralized school assignment to explore this relationship. Estimates from Denver and New York City suggest the correlation between school performance ratings and white enrollment shares reflects selection bias rather than causal school value-added. In fact, value-added in these two cities is essentially unrelated to white enrollment shares. A simple regression adjustment is shown to yield school ratings that are uncorrelated with race, while predicting value-added as well or better than the corresponding unadjusted measures.

Joshua Angrist
Department of Economics, E52-436
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
angrist@mit.edu

Peter Hull
Department of Economics
Box B, Brown University
Providence RI 02912
and NBER
peter_hull@brown.edu

Parag A. Pathak
Department of Economics, E52-426
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
ppathak@mit.edu

Christopher R. Walters
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
crwalters@econ.berkeley.edu

1 Introduction

In the Fall of 2021, US News and World Report released long-anticipated rankings of American middle and elementary schools, based on test scores and other measures of student achievement. These and other school ratings—such as those of GreatSchools.org, Niche.com, and various state accountability offices—meet a growing demand for information on school quality. The intense public interest in school performance is also clear on real estate sites like Zillow and Redfin, which feature school ratings prominently. School ratings affect families' choices of where to live and where to enroll (Bergman and Hill, 2018; Hasan and Kumar, 2019), as well as district decisions to restructure schools (Rockoff and Turner, 2010; Abdulkadiroğlu et al., 2016; Cohodes et al., 2021).

Do school ratings serve the public interest? Journalists like Barnum and LeMee (2019) have focused attention on the strong correlation between widely reported rankings and the racial make-up of schools. In urban districts enrolling large numbers of non-white students, highly-rated schools tend to enroll disproportionate shares of white and Asian students. For example, enrollment at US News' top five New York City middle schools is 80 percent white and Asian, compared with the 35 percent white and Asian share in the district as a whole.[1] Statistics like these suggest that links between published school ratings and racial composition may contribute to racial segregation (National Fair Housing Alliance, 2006; Yoshinaga, 2016).

The correlation between school ratings and student race may reflect an uncomfortable truth: Black and white students have long attended schools of differing quality, a fact documented in economics by Welch (1973). Improvements in the quality of predominately-Black schools account for much of the reduction in Black-white wage gaps seen from the 1950s through the 1970s (Card and Krueger, 1992a,b). This progress notwithstanding, schools highly segregated (Monarrez, 2021). The higher achievement and graduation rates found at schools that enroll more white students may reflect these schools' greater impact on learning—a view reflected in decades of argument over access to selective enrollment high schools like the Boston Latin School and New York's Stuyvesant, Brooklyn Tech, and Bronx Science (Jonas, 2021).

The link between school rankings and schools' racial make-up may also be an artifact of selection bias. Higher-income and non-minority students tend to have better educational outcomes for reasons other than the quality of the schools they attend. School ratings based on student achievement levels are therefore likely to conflate school quality with the

[1] The list of top New York middle schools can be found at `https://www.usnews.com/education/k12/middle-schools/new-york`. Demographic shares are calculated for the 2018-2019 school year using the administrative data described below.

background of enrolled students. More sophisticated ratings which adjust for student demographics and lagged achievement, like conventional value-added models for teachers (e.g. Chetty et al., 2014; Rothstein, 2010, 2017) and schools (e.g. Deming, 2014; Beuermann and Jackson, 2022), may nevertheless be biased by unobserved differences in student composition. Recent research suggests such selection bias is pervasive (Abdulkadiroğlu et al., 2020). Biased rating schemes direct households to low-minority rather than high-quality schools, while penalizing schools that improve achievement for disadvantaged groups.

This paper investigates the relationship between public school ratings and student racial composition, drawing broader implications for school assessment systems. Our analysis focuses on two properties of ratings: *predictive accuracy*, defined as the rating's r-squared in a regression of a school's true causal effect on achievement, and *racial imbalance*, defined as the slope in a regression of school ratings on white enrollment shares. If schools that enroll more white students tend to be better, in the sense of having higher causal value-added, those wishing to inform the public about school quality face an uncomfortable trade-off between predictive accuracy and racial imbalance.

Our findings show this trade-off to be much smaller than the correlation between school ratings and racial composition suggests. This conclusion is reached in two steps. First, we derive a simple but novel characterization of the theoretical link between accuracy and imbalance, based on unobserved school quality. Second, we estimate the components of this trade-off using the random variation in school attendance generated by centralized school assignment systems (Abdulkadiroğlu et al., 2017, 2022). Specifically, we adapt the instrumental variables value-added model (IV VAM) approach of Angrist et al. (2021) to gauge relationships between causal value-added, student race, and school ratings.

We study the trade-off between predictive accuracy and racial imbalance for New York City and Denver middle schools. Both districts allocate seats using a centralized match that generates partially randomized variation in school assignment, supplying the instruments needed for IV VAM. These two districts are also central to discussions of segregation and school access: New York is America's largest district, with a long history of de facto segregation, while Denver is a majority Hispanic district with a unified enrollment match combining charter and traditional public schools.

School performance ratings based on achievement levels and on achievement growth are both highly correlated with schools' racial composition in New York and Denver. Our analysis suggests this correlation is largely an artifact of selection bias. IV VAM estimates show causal value-added is unrelated to racial composition in both cities. Together, these findings imply that a conventional progress-based rating adjusted to be uncorrelated with student race has predictive accuracy slightly better than that of the corresponding unadjusted measure.

Moreover, in both New York and Denver, this racially-balanced progress rating essentially coincides with an optimal rating constructed to best predict causal value-added as a function of conventional progress ratings, student race, and school sector. Racially-balanced ratings may thus represent a rare “free-lunch” for school accountability policy: a simple adjustment to existing ratings, requiring only data on student race, eliminates racial imbalance while also improving the ratings’ value as predictors of true school quality.

2 Settings and Data

The Denver sample includes students applying for sixth-grade seats at any middle school in the Denver Public Schools district between the 2012-2013 and 2018-2019 school years. The New York sample includes sixth-grade applicants to New York City middle schools for the 2016-2017 through 2018-2019 school years. Data include the school preferences and priorities submitted by each applicant and the assignments generated by each district’s centralized school assignment system. We also have data on subsequent school enrollment, student demographics, and achievement scores.[2] Denver outcomes are from the Colorado Student Assessment Program (CSAP) and Colorado Measures of Academic Success (CMAS) standardized tests. New York outcomes come from New York state achievement tests. The main outcome for our analysis combines scaled math and ELA scores in sixth grade, standardized to be mean zero and standard deviation one in each city and year. Our combined math and ELA measures are similar to ratings reported by GreatSchools.org, school districts, and states.

Students in Denver rank up to five schools in the district. Admissions priorities are based on criteria like sibling status and the applicant’s residential neighborhood. The deferred acceptance (DA) algorithm with a single lottery tie-breaker assigns students to schools. New York school applicants rank up to 12 academic programs; while schools may host more than one program, our analysis aggregates multiple programs to the school level. The New York DA algorithm features a variety of tie-breakers, with “unscreened” schools using a common random lottery number and “screened” schools using non-random tie-breakers such as past test scores and grades.

Our empirical strategy leverages the randomness embedded in each city’s school assignment mechanism. We follow Abdulkadiroğlu et al. (2017, 2022) in computing each applicant’s risk (i.e. probability) of assignment to each school as a function of the applicant’s school preferences and priorities. Assignment risk for Denver applicants is computed using the propensity score formula derived by Abdulkadiroğlu et al. (2017). This formula is an

[2] The samples analyzed here are derived from those used in Angrist et al. (2021).

analytical large-market approximation to the school assignment probability for DA with a lottery tie-breaker.[3] Assignment risk for New York applicants is computed as described in Abdulkadiroğlu et al. (2022). New York assignment risk depends in part on bandwidths for screened school tie-breakers, similar to those used in standard regression discontinuity designs.[4] Score conditioning yields a stratified randomized trial. Conditional on assignment risk, school assignment is independent of applicant characteristics, both observed and unobserved—an application of the Rosenbaum and Rubin (1983) propensity score theorem.

Our analysis focuses on two achievement-based measures of school quality that replicate widely-disseminated state ratings for Colorado and New York State. *Levels* ratings consist of the share of students scored as proficient on state assessments, averaged across math and English language arts (ELA) tests. *Progress* ratings are based on year-to-year improvement in the average math and ELA achievement percentiles of enrolled students. This mirrors the student growth percentiles reported by many states and districts, as well as the GreatSchools.org Student Progress Rating. Our interest in progress ratings is partly motivated by previous findings that growth-type measures more accurately predict school quality (Angrist et al., 2017, 2021). Ratings are computed separately for every school and year, and are standardized to be mean zero with a standard deviation matching our estimated standard deviation of school value-added, detailed below. Appendix B.1 details the school ratings construction.

Appendix Table A1 summarizes the students and schools in the Denver and New York samples, separately for all enrolled students and for the subsample of applicants for whom school assignment has a random component. We refer to the latter group as the sample with risk.[5] As is typical in large urban districts, most Denver and New York students are from disadvantaged backgrounds, with over 70 percent eligible for a subsidized lunch. In both districts the demographic characteristics, enrollment status, and baseline scores of applicants with assignment risk are similar to those of the full sample of sixth-grade students. The New York sample includes 1,584 school-year observations with a median enrollment of 83 students. The Denver sample includes 435 school-year observations with a median enrollment of 81 students.[6]

The natural experiment induced by centralized assignment is validated by 1, which compares the characteristics of students offered seats at higher-rated and lower-rated schools

[3] The Denver score is computed using the *formula score* described in Abdulkadiroğlu et al. (2017).

[4] The New York score is the *local DA score* described in Section 4.2 of Abdulkadiroğlu et al. (2022). Bandwidths used here are computed as suggested by Calonico et al. (2019).

[5] Formally, applicants in this sample (indexed by i) have a propensity score p_{ij} strictly between zero and one for at least one school j. Roughly a quarter of the students in each sample face some assignment risk.

[6] The 10th (90th) percentile of school-year enrollment is 36 (279) in New York and 19 (141) in Denver. Appendix Table A2 further summarizes the samples of schools in both settings.

(these comparisons are based on the progress rating). Uncontrolled comparisons show large differences in characteristics between those offered seats at high- and low-rated schools, but these differences vanish when adjusted for assignment risk. The fact that risk adjustment balances observed characteristics suggests unobserved characteristics are likely balanced as well.[7]

Levels and progress ratings are highly correlated with the racial composition of schools, a fact documented in Figure 1. Specifically, the figure plots average school ratings computed conditional on share white in bins of width 0.1, along with the corresponding regression line fit to school-level data. Evidence of racial imbalance is especially strong for levels ratings. In New York, a regression of levels ratings on share white yields a slope coefficient of 0.70 with a robust standard error of 0.03. The standard deviation of each rating equals roughly 0.2, so this coefficient implies that a ten percentage-point increase in share white is associated with a rating increase of about 0.35 standard deviations. The corresponding slope falls to 0.22 for progress, but the relationship remains clear and statistically precise. Evidence of racial imbalance for Denver is similar, with coefficients of 0.85 for levels and 0.38 for progress (both precisely estimated).

3 Econometric Framework

The distinction between causal value-added and selection bias is cast here in terms of a constant-effects causal model of education production. Consider a population of students, each attending one of J schools in a district. Student i's potential academic achievement at school $j \in \{1, ..., J\}$, denoted Y_{ij}, is:

$$Y_{ij} = \beta_j + \varepsilon_i, \tag{1}$$

where β_j gives the contribution of attendance at school j to achievement. Parameter β_j is school j's *quality* or *value-added*. Random variable ε_i reflects other factors that influence a student's academic achievement, such as family background, motivation, and ability.

Equation (1) is a constant-effects model because ε_i is assumed to vary across students but not schools. For any two schools, j and k, and any applicant, i, $Y_{ij} - Y_{ik} = \beta_j - \beta_k$ gives the causal effect of attending j rather than k. This constant-effects setup allows us to focus

[7] Balance checks regress student characteristics on the progress rating of the school where applicants are offered a seat, along with a dummy indicating whether the applicant was offered a seat anywhere. Risk controls consist of the expected progress rating and the probability of receiving any offer. The former is computed as a score-weighted average of the school quality measure, following Borusyak and Hull (Forthcoming). Appendix Table A3 shows that follow-up rates for key outcomes are unrelated to assigned school ratings, conditional on assignment risk.

on selection bias rather than treatment effect heterogeneity.[8]

The outcome observed for student i, denoted Y_i, equals the potential outcome associated with the school he or she attends. Letting D_{ij} be an indicator for student i's enrollment at school j, we have:

$$Y_i = \sum_j Y_{ij} D_{ij} = \sum_j \beta_j D_{ij} + \varepsilon_i. \tag{2}$$

The average outcome at school j is given by $E[Y_i|D_{ij} = 1]$. School attendance is not randomly assigned, so these average outcomes may be a poor guide to causal effects. In particular, for any school j, $E[Y_i|D_{ij} = 1] = \beta_j + E[\varepsilon_i|D_{ij} = 1]$ which differs from β_j when schools are chosen based on factors that are correlated with ε_i.

Schools are also distinguished by the demographic composition of their student bodies. Let W_j denote the share of students enrolled in school j designated as white, i.e., $W_j = E[w_i \mid D_{ij} = 1]$, where w_i indicates student i's race. Correlation between share white and school ratings may arise because of a relationship between W_j and β_j, in which case the rating accurately reveals a demographic gap in school quality. Alternatively, this correlation may arise at least in part because D_{ij} is correlated with (w_i, ε_i): a case of selection bias.

3.1 Racial Imbalance and Predictive Accuracy

Because β_j is unobserved, educational authorities report an imperfect rating, R_j, computed as a function of student achievement. As in earlier work on value-added (e.g., Angrist et al., 2016, 2017), we treat school-level characteristics—here ratings, quality, and share white—as random variables. Our investigation of the relationship between school ratings and racial composition focuses on two aspects of the distribution of school ratings:

Definition. The *predictive accuracy* of school rating R_j is defined as $\rho_R = \frac{Cov(\beta_j, R_j)^2}{Var(\beta_j)Var(R_j)}$. The *racial imbalance* of school rating R_j is given by $\mathcal{I}_R = \frac{Cov(W_j, R_j)}{Var(W_j)}$.

The predictive accuracy of a rating scheme is the r-squared from a regression of school quality on ratings. Parents or policymakers seeking to identify effective schools should prefer ratings with higher ρ_R. A rating scheme's racial imbalance is the slope coefficient from a regression of R_j on W_j. These features are defined for any choice of R_j, including β_j itself, so $\mathcal{I}_\beta$ denotes the slope coefficient from a regression of β_j on W_j.[9]

[8] Angrist et al. (2017, 2021) find little evidence of effect heterogeneity in lottery-based analyses of school value-added in the cities studied here. This conclusion is supported by estimates that allow school effects to vary with student characteristics.

[9] In practice the school quality distributions we study, like school ratings, are year-specific. See Appendix B.1 for details.

Racially imbalanced rating schemes may favor schools with a higher share white regardless of school quality. To ameliorate this, race-balanced ratings can be constructed as the residual from a regression of R_j on W_j:

$$R_j = \gamma + \lambda W_j + \tilde{R}_j., \tag{3}$$

where $\lambda = \mathcal{I}_R$. By construction, $\tilde{R}_j$ is uncorrelated with W_j and thus has racial imbalance $\mathcal{I}_{\tilde{R}} = 0$.

Although racial imbalance is easily eliminated, this may come at the cost of reduced predictive accuracy. To describe this trade-off, consider first the coefficients on ratings in the following two predictive regressions for school quality:

$$\beta_j = \mu + \varphi R_j + \nu_j, \tag{4}$$

$$\beta_j = \tilde{\mu} + \tilde{\varphi} R_j + \tau W_j + \tilde{\nu}_j. \tag{5}$$

Predictive accuracy is the r-squared for (4), and is therefore proportional to φ^2, while $\tilde{\varphi}$ coincides with the coefficient from a regression of β_j on the ratings residual $\tilde{R}_j$. We refer to φ and $\tilde{\varphi}$ as *forecast coefficients*, quantifying the relationship between school quality and imperfect ratings.

Suppose that schools with a higher share of white students tend to be rated higher, as in Figure 1: i.e. $\mathcal{I}_R > 0$. The two forecast coefficients are then related as follows:

Proposition 1. *Suppose $\mathcal{I}_R > 0$. Then, $\tilde{\varphi} > \varphi$ if and only if $\tau < 0$.*

Proof. By the omitted variables bias formula, $\varphi = \tilde{\varphi} + \tau \frac{Cov(R_j,W_j)}{Var(R_j)}$. So, $\tilde{\varphi} > \varphi$ if and only if $\tau < 0$ when $cov(R_j, W_j) > 0$. □

Proposition 1 shows that, given the gradient in Figure 1, race-adjusted ratings generate a larger forecast coefficient whenever the coefficient on share white in the long forecast regression (5) is negative. This happens in a scenario in which schools with a higher share white tend to have value-added below that of other schools with the same rating. This pattern arises, for example, with a rating scheme that rewards share white in a school system where race predicts ε_i but not school quality.

The effect of racial balancing on predictive accuracy is determined by the ratio of the forecast coefficients defined by (4) and (5), along with τ and the racial imbalance in school quality:

Proposition 2. *Suppose $\mathcal{I}_R > 0$ and $\tilde{\varphi} > 0$. Then $\rho_{\tilde{R}} > \rho_R$ if and only if $\mathcal{I}_\beta < -\tau(\varphi/\tilde{\varphi})$.*

Proof. See Appendix A. □

This result is especially sharp in a scenario where school quality is unrelated to race, so $\mathcal{I}_\beta = 0$. In this case, if ratings are racially imbalanced ($\mathcal{I}_R > 0$) but still informative, then $\tau < 0$ and $\rho_{\tilde{R}} > \rho_R$.[10] More generally, Proposition 2 shows that when τ is negative racial adjustment increases the predictive value of ratings as long as race is a sufficiently weak predictor of school quality. In this case, Proposition 2 shows that racial adjustment offers a free lunch, boosting predictive accuracy by eliminating racial imbalance.

An analyst interested in maximizing predictive accuracy might combine information on racial make-up with ratings data using fitted values from (5):

$$\beta_j^* = \tilde{\mu} + \tilde{\varphi} R_j + \tau W_j. \tag{6}$$

This best linear predictor of school quality may improve and cannot reduce predictive accuracy relative to both R_j and $\tilde{R}_j$, since the extra regressor, W_j, cannot reduce r-squared.[11] The question of whether β_j^* mitigates racial imbalance is addressed by the following result:

Proposition 3. *The racial imbalance of the fitted values from regression* (5) *and the racial imbalance of causal value-added coincide:* $\mathcal{I}_{\beta^*} = \mathcal{I}_\beta$.

Proof. $Cov(W_j, \tilde{\nu}_j) = 0$, so $\frac{Cov(W_j,\beta_j)}{Var(W_j)} = \frac{Cov(W_j,\beta_j^*+\tilde{\nu}_j)}{Var(W_j)} = \frac{Cov(W_j,\beta_j^*)}{Var(W_j)}$. □

This result formalizes the intuition that any racial imbalance in school quality is captured by the coefficient on W_j in the model generating β_j^*.

In summary, Propositions 1-3 show that the trade-off between the predictive power and racial imbalance of forecast coefficients φ and $\tilde{\varphi}$, the coefficient τ in equation (5), and the racial imbalance of value added, $\mathcal{I}_\beta$. The challenge in applying these results is that school quality, β_j, is unobserved. To surmount this challenge, we estimate the determinants of predictive accuracy and racial imbalance for alternative ratings using the IV VAM empirical strategy detailed in Angrist et al. (2021). Specifically, we use instruments to estimate forecast parameters $\varphi, \tilde{\varphi}$, and τ. IV VAM also yields a measure of $\mathcal{I}_\beta$, the slope from a regression of school quality on share white, and an estimate of the total variance of β_j, needed to calculate the predictive accuracy of each rating.

[10] If $\mathcal{I}_\beta = 0$, then τ is proportional to $Cov(\beta_j, W_j - \alpha R_j) = -\alpha Cov(\beta_j, R_j)$ where α is the coefficient from a regression of W_j on R_j. When $\mathcal{I}_R > 0$, $\alpha > 0$, so $\tau < 0$ when $\varphi \propto Cov(\beta_j, R_j) > 0$.

[11] To see this for $\tilde{R}_j$, let $\hat{R}_j$ be the fitted values from (3) and write equation (6) as

$$\beta_j^* = \tilde{\mu} + \tilde{\varphi}\hat{R}_j + (\tilde{\varphi}\tilde{R}_j + \tau W_j).$$

The term in parentheses on the right-hand side is orthogonal to the balanced rating, $\tilde{R}_j$, so the variance of β_j^* exceeds the variance of $\tilde{R}_j$.

3.2 Identification and Estimation

IV VAM starts with an augmented version of regression (5) that incorporates additional predictors of school quality. The augmented model can be written:

$$\beta_j = M_j'\psi + \xi_j, \tag{7}$$

where M_j denotes a vector of quality predictors. M_j includes a constant, school ratings, share white, and school sector dummies. Forecast regression (7) is a linear projection, so $E[M_j\xi_j] = 0$ by definition of forecast residual ξ_j. Substituting this projection into the causal model (2) yields:

$$\begin{aligned} Y_i &= \sum_j (M_j'\psi + \xi_j)D_{ij} + \varepsilon_i \\ &= M_{j(i)}'\psi + \xi_{j(i)} + \varepsilon_i, \end{aligned} \tag{8}$$

where $M_{j(i)} = \sum_j M_j D_{ij}$ and $\xi_{j(i)} = \sum_j \xi_j D_{ij}$ denote the school characteristics and forecast residual for student i's school, indexed by $j(i)$. Equation (7) is a linear projection, but equation (8) need not be since elements of $M_{j(i)}$ are correlated with ε_i. IV VAM therefore uses centralized school assignment offers, denoted Z_{ij} for school j, as instruments for the school characteristics in $M_{j(i)}$.[12]

The IV VAM estimating equation includes a vector of individual-level control variables, X_i, including school assignment risk and other applicant characteristics. Controlling for the latter isn't necessary for identification, but may boost precision.[13] Let θ denote the coefficient from a regression of the composite residual $\xi_{j(i)} + \varepsilon_i$ on X_i, with associated residual η_i. The IV VAM estimating equation can then be written

$$Y_i = M_{j(i)}'\psi + X_i'\theta + \eta_i, \tag{9}$$

where $E[X_i\eta_i] = 0$ by definition of θ.

The addition of risk controls to the covariate vector in a linear model is sufficient to ensure offer instruments Z_{ij} are uncorrelated with unobserved applicant background and ability, ε_i. Importantly, however, residual η_i in (9) depends on a school component, $\xi_{j(i)}$, as well as ε_i.

[12] An alternative instruments school enrollment indicators in equation (2), thereby estimating β_j directly. This is infeasible here, however, because some schools are undersubscribed. IV VAM addresses the identification problem arising from the fact that we have fewer instruments than schools.

[13] Additional controls are functions of 5th grade math and ELA scores, the demographic variables listed in Appendix Table A1, and year fixed effects interacted with lagged scores and demographic characteristics. Risk controls for New York include local linear functions of the relevant screened-school tie-breakers; see Abdulkadiroğlu et al. (2022) for details.

The former reflects determinants of value added not explained by M_j and can be thought of as arising from violations of the IV exclusion restriction that underpins identification in this context. Angrist et al. (2021) formulates sufficient conditions for IV VAM estimates to be consistent in the face of such violations. These conditions require the relationship between individual school offers and residual school quality to average to zero over schools.

The IV VAM exclusion restriction is made more plausible by including strong predictors of school quality in M_j. Such mediators reduce and perhaps even eliminate variation in residual school quality, ξ_j. In our implementation, M_j includes the levels and progress ratings, share white, a dummy for charter schools (in Denver), and a dummy for screened schools (in New York). By instrumenting average test score levels and growth measures, we avoid mechanical biases from simply regressing outcomes on outcome averages.

The parameters in (9) are estimated by two-stage least squares (2SLS). This yields estimates of ψ in equation (7), defined as the regression of β_j on the full vector of school characteristics, M_j. Coefficients in shorter projections of β_j on subsets of M_j can then be generated by application of the omitted variables bias formula. For example, the coefficients in (5) are obtained from a partition such that $M_j = (M'_{1j}, M'_{2j})'$, with $M_{1j} = (1, R_j, W_j)'$ and $\psi = (\psi'_1, \psi'_2)'$ partitioned correspondingly. We then have:

$$(\tilde{\mu}, \tilde{\varphi}, \tau)' = \psi_1 + E[M_{1j}M'_{1j}]^{-1}E[M_{1j}M'_{2j}]\psi_2. \tag{10}$$

This two-step approach uses 2SLS estimates of (9) as the common foundation for forecast regressions of any shorter length. As a by-product, the minimized 2SLS minimand (an overidentification test statistic) generates a quadratic form proportional to the variance of β_j. This variance is used in the formula for predictive accuracy.[14]

4 Results

School quality is unrelated to the share of enrolled students who are white in the sample of New York schools. This can be seen in the first column of Panel A in Table 2, which reports estimates of the projection of β_j on share white and a screened school indicator for

[14] Specifically, the variance of ξ_j is estimated by $\frac{(Y-Q\hat{\phi})'P_{\tilde{Z}}(Y-Q\hat{\phi})}{tr(\hat{\Pi}Z'Z\hat{\Pi})}$, where Y is the vector of outcomes, Q is the matrix of variables to be instrumented including covariates, $P_{\tilde{Z}}$ is the projection matrix for the instruments $\tilde{Z}$ after partialling out covariates, $\hat{\phi}$ is the vector of 2SLS coefficient estimates, and $\hat{\Pi}$ is the matrix of first-stage coefficient estimates. Appendix I of Angrist et al. (2021) derives this formula. The version used here omits bias-correction terms which yield qualitatively similar results; see Angrist et al. (2021) for details.

schools in New York.[15] The full set of IV VAM estimates underlying these results appears in Appendix Table A6.[16] Both of the derived coefficient estimates in column 1 of Table 2 are small and significantly insignificant. The share white coefficient is estimated precisely enough to rule out a racial imbalance as large as 0.124 on the basis of 95 percent confidence interval coverage. The large racial imbalance estimate of 0.687 in column 3, by contrast, shows that share white is highly predictive of school ratings based on test score levels—as seen in Figure 1. Together, the results in columns 1 and 3 imply that the strong relationship between school ratings and share white in New York reflects selection bias rather than school quality.[17]

Levels ratings are weakly related to school quality in New York: the estimated forecast coefficient in column 2 of Table 2 (Panel A) shows that a one standard deviation improvement in test score levels is associated with a 0.21 standard deviation increase in causal value-added.[18] Column 4 reports estimates of $\tilde{\varphi}$ and τ in forecast equation (5), computed by adding share white and screened-school status to ratings as predictors of school quality. Estimated coefficients on the screened-school dummy and share white are both negative and significantly different from zero. This matches the pattern discussed above, wherein schools that enroll more white students, as well as highly sought-after screened schools, are of lower quality than other similarly-rated schools.

Column 5 of Table 2 shows progress ratings predict New York school quality with a forecast coefficient of about 0.76—a marked improvement relative to the levels rating. But progress ratings are compromised by selection bias too. Column 6 in Panel A reports an estimated share white coefficient of 0.22 in a regression of progress which controls for a screened-school dummy. Column 7 shows the progress coefficient remains high when quality is predicted by progress and share white, but share white is again negatively related to quality. Like the estimates in column 4, this pattern reflects the fact that quality and share white are

[15] Appendix Tables A4 and A5 report results from models that replace share white with share white or Asian and with the share of students not eligible for a free or reduced price lunch (FRPL). These variations yield results similar to those in Table 2.

[16] The first-stage F-statistics for these estimates, computed as Kleibergen-Paap (2006) robust Wald test statistics, are above the rule-of-thumb threshold of 10 commonly used to diagnose weak instrument bias. The 2SLS estimates in the table are also close to just-identified IV estimates reported in Table A6, from models where weak instrument bias is unlikely to be a concern. This just-identified estimator replaces individual school offer dummies as instruments with values of the mediator at the offered school, one for each mediator. Over-identified limited information maximum likelihood and the bias-corrected IV estimator in Kolesár et al. (2015) are likewise similar to the 2SLS estimates reported here.

[17] Appendix Table A7 tests the equality of IV estimates of racial imbalance in school quality and OLS estimates of the racial imbalance in ratings. These tests reject decisively in New York.

[18] As detailed in Appendix B.1, each rating is scaled to have the same standard deviation as estimated for value-added, so that the forecast coefficient can be interpreted as the standard deviation gain in causal value-added associated with a one standard deviation increase in the rating.

unrelated, so that disproportionately white and screened schools are, on average, over-rated. The fact that progress ratings exhibit modest selection bias, while improving markedly over the predictive accuracy of the levels rating, is consistent with past findings on bias in school value-added models (Angrist et al., 2017, 2021). The fact that racial imbalance decreases for more accurate ratings reflects the finding that school quality is essentially uncorrelated with race.

Analogous results for Denver, reported in Panel B of Table 2, are qualitatively similar to those for New York, though these smaller-district estimates are less precise. Column 1 shows a statistically insignificant relationship between school quality and share white, while Denver's many charter schools generate a precisely estimated achievement gain of about 0.10 standard deviations. As in New York, share white predicts levels more than progress (compare columns 3 and 6 in Panel B), but both predictive relationships for ratings are strong. Also as in New York, multivariate quality projections for Denver yield negative (though more imprecise) estimated coefficients on share white when ratings are included as an explanatory variable; see columns 4 and 7 of Panel B.[19]

Appendix Figure A1 highlights implications of the results in Table 2 by plotting alternative ratings against share white in New York and Denver. The figure shows the estimated conditional expectation function (CEF) for three ratings, computing in 10-point bins, along with a regression fit to the underlying school-level data. As in Figure 1, the relationship between the progress rating and share white for New York schools is positive. Race-balanced progress, computed as the residual from a regression of progress on share white, generates a flat regression fit by construction. The best linear predictor of New York school quality given the progress rating, share white, and screened school status (the fitted value from the model generating column 7 of Table 2) generates a similar pattern.

IV VAM estimates suggest ratings for Denver are less compromised by selection bias than the corresponding estimates for New York, with larger forecast coefficients for both levels and progress. Share white is also more strongly predictive of progress ratings in Denver than in New York (compare the estimates for the two cities in column 6 of Table 2). Consistent with these estimates, the CEF for the best linear predictor of Denver school quality plotted in Panel B of Appendix Figure A1 is weakly dependent on share white. Even so, the best linear predictor for Denver school quality rises much less steeply in share white compared to the CEF of the raw progress rating.

Table 3 summarizes our investigation with estimates of predictive accuracy and racial imbalance for alternative ratings. In both New York and Denver, progress ratings are far

[19] Denver estimates too imprecise to rule out moderate degrees of racial imbalance in causal value added. As for New York, however, ratings are significantly more imbalanced than school quality.

more accurate than levels ratings while also being much more weakly correlated with share white. This improvement notwithstanding, progress remains substantially correlated with race. Race-balanced ratings boost predictive accuracy in both cities. The best linear predictor of school quality given progress ratings, share white, and a sector dummy has predictive accuracy only slightly better than that of race-balanced progress. This is explained by the fact that the best linear predictor of school quality depends little, if at all, on race.

5 Conclusions

This paper uses the random assignment embedded in centralized school assignment mechanisms to study the relationship between school ratings, school quality, and student race. In Denver and New York middle schools, the fact that schools with more white students are highly rated reflects selection bias rather than educational quality. As a result, ratings purged of correlation with race predict school quality as well or better than standard measures.[20]

Denver and New York are just two districts, of course; but the fact that our analysis yields similar conclusions in both is noteworthy. Denver enrolls many more Hispanic students and runs a unified admissions system that includes charter schools. It's also worth noting that the correlation between race and accountability measures documented in Figure 1 is visible in districts nationwide. Across all US schools in 2018, a regression of GreatSchools' levels school ratings on share white yields a coefficient of 0.632 while the corresponding regression for the GreatSchools' progress measure is only 0.310 (see Appendix Table A8; these regressions control for district fixed effects and charter status).[21] Larger differences in correlation appear in New York State and Colorado, the states containing our study districts. Thus estimates for many schools beyond those in New York City and Denver are consistent with our claim that the association between race and achievement levels in the typical urban district is primarily due to selection bias. The growing importance of centralized assignment should allow more rigorous validation of this claim in the not-too-distant future.[22] An equally important question for future work, requiring empirical methods distinct from those used here, is whether our findings extend to racial imbalance *across* districts.

[20] Other efforts in this direction, inspired by similar concerns with possibly misleading racial imbalance, include the GreatSchools Equity Rating (`https://www.greatschools.org/gk/ratings-methodology/#methodology-equity-rating`).

[21] Levels is GreatSchools' Test Score Rating, and progress is GreatSchools' Student Progress Rating when available and Academic Progress Rating otherwise.

[22] A review of enrollment portals in the 100 largest districts shows more than a third of urban students attend schools in districts that assign seats centrally, while 83% attend schools in districts with at least some random assignment. See Appendix Table A9 for these statistics.

Our analysis leaves open the question of how racially-balanced school ratings might affect household decision-making. Households appear to respond to school performance ratings (Hastings and Weinstein, 2008; Bergman and Hill, 2018; Bergman et al., 2020; Houston and Henig, 2023; Campos and Kearns, 2023). Credible racially-balanced quality information may therefore increase the demand for high-quality schools with lower white enrollment. At the same time, school choice may respond more to peer characteristics than to value-added (Rothstein, 2006; Abdulkadiroğlu et al., 2020). We hope to study the extent to which households respond to improved measures of school quality in future work.

A Appendix Proof of Proposition 2

Predictive accuracy for R_j and $\tilde{R}_j$ is given by $\rho_R = \frac{\varphi^2 Var(R_j)}{Var(\beta_j)}$ and

$$\rho_{\tilde{R}} = \frac{\tilde{\varphi}^2 Var(\tilde{R}_j)}{Var(\beta_j)} = \frac{\tilde{\varphi}^2 \left(Var(R_j) - \lambda^2 Var(W_j)\right)}{Var(\beta_j)},$$

respectively, where the latter expression uses fact that the fitted values and residuals in regression (3) are uncorrelated. The change in r-squared after residualizing is therefore proportional to

$$\begin{aligned}
(\rho_{\tilde{R}} - \rho_R) Var(\beta_j) &= \tilde{\varphi}^2 \big(Var(R_j) - \lambda^2 Var(W_j)\big) - \varphi^2 Var(R_j) \\
&= (\tilde{\varphi} - \varphi)(\tilde{\varphi} + \varphi) Var(R_j) - \tilde{\varphi}^2 \lambda^2 Var(W_j) \\
&= -\tau\lambda \frac{Var(W_j)}{Var(R_j)} (\tilde{\varphi} + \varphi) Var(R_j) - \tilde{\varphi}^2 \lambda^2 Var(W_j) \\
&= -\big(\tau(\tilde{\varphi} + \varphi) + \tilde{\varphi}^2 \lambda\big) \lambda Var(W_j), \qquad (11)
\end{aligned}$$

using the fact that $\tilde{\varphi} - \varphi = -\tau\lambda \frac{Var(W_j)}{Var(R_j)}$ by the proof to Proposition 1 and the definition of $\lambda = \frac{Cov(W_j, R_j)}{Var(W_j)}$. Since $\lambda = \mathcal{I}_R$ by definition and $\mathcal{I}_R > 0$, equation (11) shows that when $\tilde{\varphi} > 0$, $\rho_{\tilde{R}} > \rho_R$ if and only if

$$\tau + \tilde{\varphi}\lambda < -\tau \frac{\varphi}{\tilde{\varphi}}. \qquad (12)$$

By the omitted variables bias formula $\tau + \tilde{\varphi}\lambda = \mathcal{I}_\beta$, completing the proof. □

References

Abdulkadiroğlu, A., J. D. Angrist, P. D. Hull, and P. A. Pathak (2016): "Charters Without Lotteries: Testing Takeovers in New Orleans and Boston," *American Economic Review*, 106(7), 1878–1920.

Abdulkadiroğlu, A., J. D. Angrist, Y. Narita, and P. A. Pathak (2017): "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation," *Econometrica*, 85, 1373–1432.

——— (2022): "Breaking Ties: Regression Discontinuity Design Meets Market Design," *Econometrica*, 90, 117–151.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020): "Do Parents Value School Effectiveness?" *American Economic Review*, 110, 1502–39.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2016): "Interpreting Tests of School VAM Validity," *American Economic Review: Papers & Proceedings*, 106, 388–392.

——— (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, 132, 871–919.

——— (2021): "Credible School Value-Added with Undersubscribed School Lotteries," MIT Blueprint Labs Working Paper. Forthcoming, *Review of Economics and Statistics*.

Barnum, M. and G. L. LeMee (2019): "Looking For a Home? You've Seen GreatSchools Ratings. Here's How They Nudge Families Toward Schools With Fewer Black and Hispanic Students," Chalkbeat.

Bergman, P., E. Chan, and A. Kapor (2020): "Housing Search Frictions: Evidence from Detailed Search Data and a Field Experiment," Working Paper.

Bergman, P. and M. J. Hill (2018): "The Effects of Making Performance Information Public: Regression Discontinuity Evidence from Los Angeles Teachers," *Economics of Education Review*, 66, 104–113.

Beuermann, D. W. and C. K. Jackson (2022): "The Short and Long-Run Effects of Attending The Schools that Parents Prefer," *Journal of Human Resources*, 57, 725–746.

Borusyak, K. and P. Hull (Forthcoming): "Non-Random Exposure to Exogenous Shocks," *Econometrica*.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019): "Regression Discontinuity Designs Using Covariates," *Review of Economics and Statistics*, 101, 442–451.

Campos, C. and C. Kearns (2023): "The Impact of Public School Choice: Evidence from Los Angeles' Zones of Choice," *Quarterly Journal of Economics*.

Card, D. and A. B. Krueger (1992a): "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100, 1–40.

——— (1992b): "School Quality and Black-White Relative Earnings: A Direct Assessment," *Quarterly Journal of Economics*, 107, 151–200.

Castellano, K. E. and A. D. Ho (2013): "A Practitioner's Guide to Growth Models," *Council of Chief State School Officers*.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014): "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2563.

Cohodes, S. R., E. M. Setren, and C. R. Walters (2021): "Can Successful Schools Replicate? Scaling Up Boston's Charter School Sector," *American Economic Journal: Economic Policy*, 13, 138–67.

Colorado Department of Education (2019): "Colorado Growth Model," Fact Sheet.

Deming, D. (2014): "Using School Choice Lotteries to Test Measures of School Effectiveness," *American Economic Review: Papers & Proceedings*, 104, 406–411.

Gould, S. J. (1981): *The Mismeasure of Man*, WW Norton & Company.

GreatSchools.org (2020): "School demographics characteristics, administrative characteristics, and ratings," Data Files.

Hasan, S. and A. Kumar (2019): "Digitization and Divergence: Online School Ratings and Segregation in America," SSRN Working Paper.

Hastings, J. S. and J. M. Weinstein (2008): "Information, School Choice, and Academic Achievement: Evidence from Two Experiments," *Quarterly Journal of Economics*, 123, 1373–1414.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

Houston, D. M. and J. R. Henig (2023): "The 'Good' Schools: Academic Performance Data, School Choice, and Segregation," *AERA Open*, 9.

Jonas, M. (2021): "Are Exam Schools Really an Academic Promised Land?" Commonwealth Magazine.

Kleibergen, F. and R. Paap (2006): "Generalized Reduced Rank Tests Using the Singular Value Decomposition," *Journal of Econometrics*, 133, 97–126.

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2015): "Identification and Inference With Many Invalid Instruments," *Journal of Business & Economic Statistics*, 33, 474–484.

Monarrez, T. E. (2021): "School Attendance Boundaries and the Segregation of Schools in the US," Working paper.

National Center for Education Statistics (2021): "Digest of Education Statistics 2021 Table 215.10." Data Table.

National Fair Housing Alliance (2006): "Unequal Opportunity – Perpetuating Housing Segregation in America," Fair Housing Trends Report.

New York City Department of Education (2020): "Demographics, test scores, match priorities, and middle school admissions," Data Files.

New York State Education Department (2020): "Growth Model for Institutional Accountability 2018/19," Technical Report.

Rockoff, J. and L. J. Turner (2010): "Short-Run Impacts of Accountability on School Quality," *American Economic Journal: Economic Policy*, 2, 119–47.

Rosenbaum, P. R. and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

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

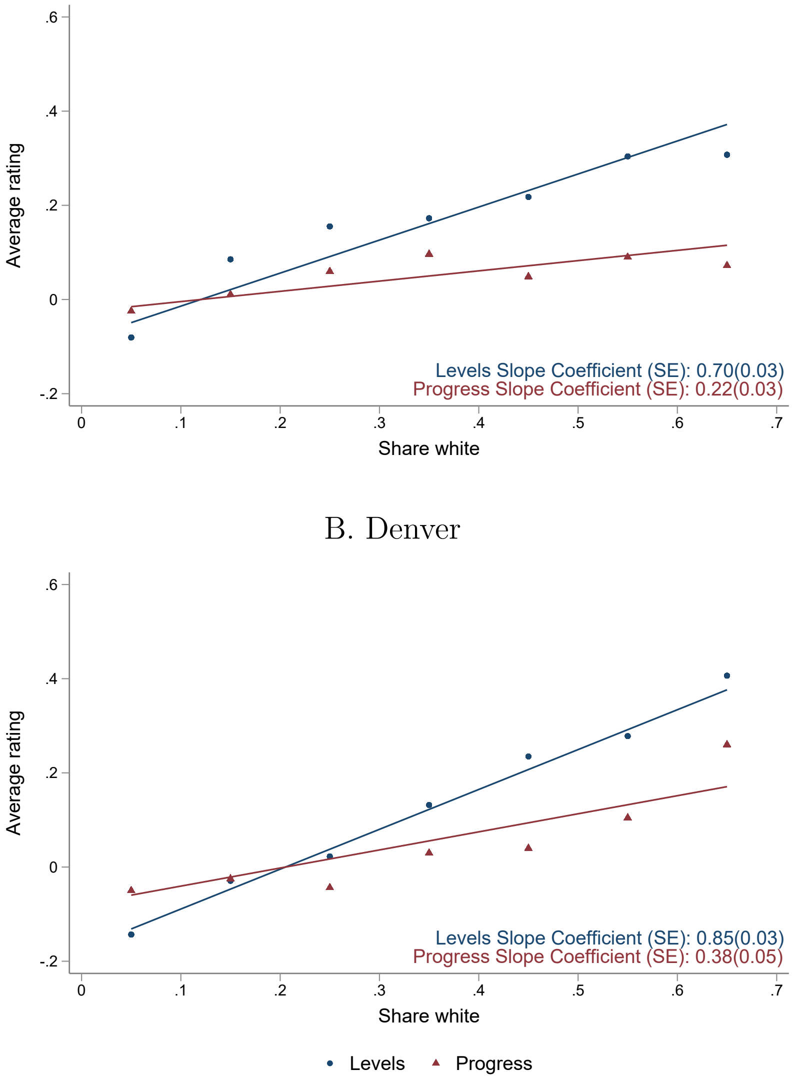
——— (2017): "Measuring the Impacts of Teachers: Comment," *American Economic Review*, 107, 1656–1684.

Rothstein, J. M. (2006): "Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions," *American Economic Review*, 96, 1333–1350.

Welch, F. (1973): "Black-White Differences in Returns to Schooling," *American Economic Review*, 63, 893–907.

Yoshinaga, K. (2016): "Race, School Ratings And Real Estate: A 'Legal Gray Area'," National Public Radio.

Figure 1. Levels, Progress, and Race



A. New York

B. Denver

Notes: Binned scatterplots depict average levels and progress ratings conditional on the share of students at a school that are white. Bins are defined by 0.1 increments in share white with the last bin grouping schools with share white ≥ 0.6. The levels rating is the mean share of students deemed proficient in math and ELA, based on sixth-grade state assessment scores. The progress rating is computed using the student growth percentile models described in Appendix B.1. Ratings are mean zero and scaled to have standard deviation equal to the standard deviation of school quality across schools in the district, roughly 0.2 in both cities.

Table 1. Statistical Tests for Balance

	New York		Denver	
	Uncontrolled	Controlled	Uncontrolled	Controlled
	(1)	(2)	(3)	(4)
Demographics				
Hispanic	-0.169	0.037	-0.481	0.033
	(0.008)	(0.025)	(0.014)	(0.045)
Black	-0.540	-0.013	0.020	0.006
	(0.007)	(0.023)	(0.010)	(0.033)
Asian	0.357	-0.030	-0.006	0.012
	(0.006)	(0.016)	(0.005)	(0.014)
White	0.360	-0.002	0.443	-0.043
	(0.005)	(0.013)	(0.013)	(0.036)
Female	0.020	0.034	-0.051	0.006
	(0.008)	(0.025)	(0.015)	(0.046)
Free/reduced price lunch	-0.274	0.045	-0.519	0.037
	(0.007)	(0.020)	(0.014)	(0.041)
Special education	-0.092	0.003	-0.054	0.000
	(0.006)	(0.020)	(0.009)	(0.027)
English language learner	0.017	0.031	-0.282	0.019
	(0.005)	(0.017)	(0.014)	(0.046)
Baseline scores				
Math (standardized)	1.02	0.020	0.858	0.090
	(0.015)	(0.046)	(0.030)	(0.092)
ELA (standardized)	0.759	-0.016	0.780	0.013
	(0.015)	(0.048)	(0.029)	(0.088)
N	184,760	46,095	37,089	8,100

Notes: This table reports balance statistics, estimated by regressing baseline covariates on the estimated progress rating of the offered school and an indicator for any offer. Rows report the estimated coefficient on the former. Estimates in columns 2 and 4 control for expected progress rating, any offer risk, and running variable controls in the New York sample. Expected progress rating is computed as a score-weighted average of the school quality measure following Borusyak and Hull (Forthcoming). Robust standard errors are reported in parentheses.

Table 2. Projections of School Quality and School Ratings on School Characteristics

		Test score levels			Test score progress		
	Value-added projection (derived)	Value-added projection (derived)	Rating projection (OLS)	Value-added projection (derived)	Value-added projection (derived)	Rating projection (OLS)	Value-added projection (derived)
Dependent variable:	School quality (β)	School quality (β)	Test score levels (R)	School quality (β)	School quality (β)	Test score progress (R)	School quality (β)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Panel A. New York				
Test score levels		0.214		0.391			
		(0.053)		(0.060)			
Test score progress					0.757		0.785
					(0.037)		(0.037)
Screened school dummy	-0.052		0.101	-0.092		-0.034	-0.025
	(0.035)		(0.014)	(0.035)		(0.016)	(0.032)
Share white	0.004		0.687	-0.265		0.222	-0.171
	(0.061)		(0.024)	(0.069)		(0.026)	(0.057)
First-stage F				23.2			
N (school-year)				1501			
			Panel B. Denver				
Test score levels		0.468		1.28			
		(0.124)		(0.207)			
Test score progress					0.859		0.975
					(0.084)		(0.099)
Charter school dummy	0.095		0.098	-0.031		0.141	-0.043
	(0.037)		(0.011)	(0.046)		(0.020)	(0.040)
Share white	0.188		0.881	-0.941		0.433	-0.235
	(0.135)		(0.027)	(0.225)		(0.051)	(0.135)
First-stage F				15.1			
N (school-year)				373			

Notes: Estimates in columns 1-2, 4-5, and 7 are from projections of school quality on the predictors listed at left. These estimates are derived from the long IV VAM coefficient estimates reported in Table A6 and computed via the omitted-variables bias formula as described in the text. Estimates in columns 3 and 6 are from models that predict ratings. These come from regressions of school ratings on share white and a school sector dummy. Robust standard errors are reported in parentheses.

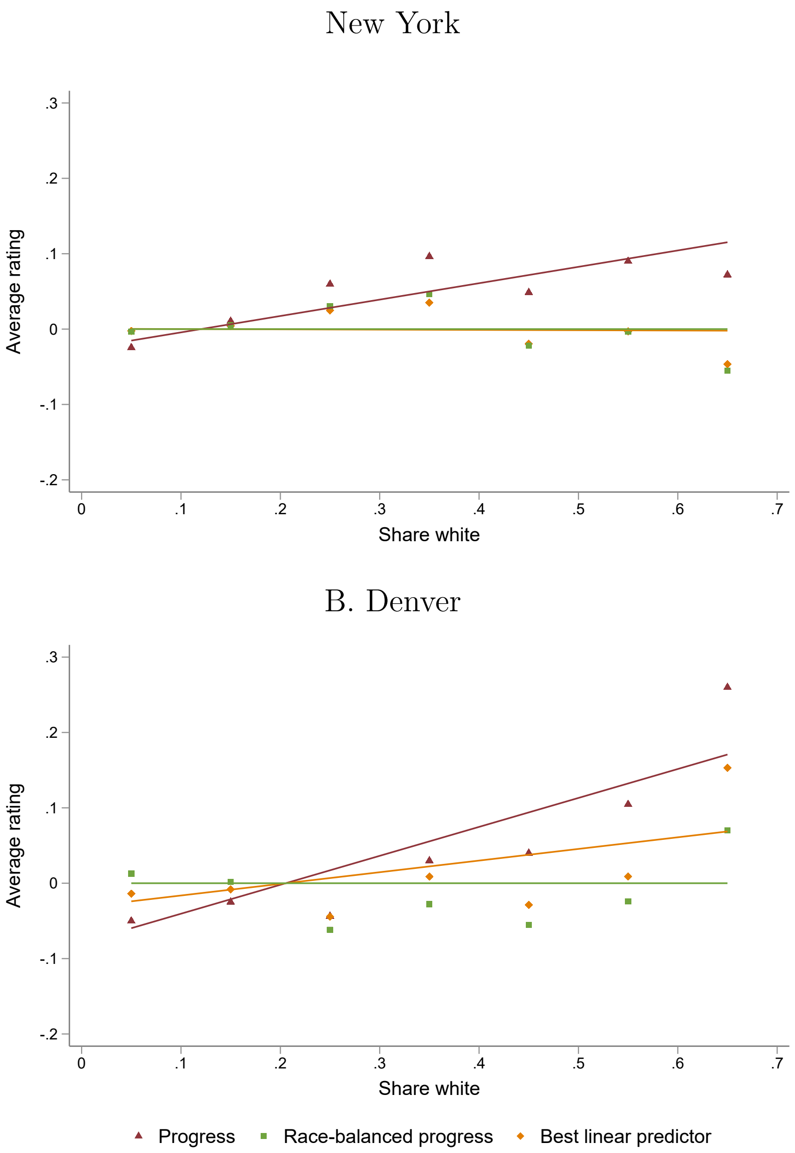
Table 3. Predictive Accuracy and Racial Imbalance

	New York		Denver	
	Predictive accuracy (ρ)	Racial imbalance $(\mathcal{I})$	Predictive accuracy (ρ)	Racial imbalance $(\mathcal{I})$
	(1)	(2)	(3)	(4)
1. Test score levels	0.046	0.702	0.219	0.846
		(0.026)		(0.027)
2. Test score progress	0.573	0.217	0.738	0.384
		(0.026)		(0.050)
3. Race-balanced progress	0.596	0.000	0.751	0.000
		-		-
4. Best linear predictor	0.598	-0.004	0.783	0.154
		(0.061)		(0.134)

Notes: This table reports predictive accuracy (ρ_R) and racial imbalance ($\mathcal{I}_R$) for alternative school ratings. Predictive accuracy is derived from IV VAM regressions of causal school quality on ratings. In rows 1-2 and 4, racial imbalance is the bivariate coefficient from a regression of ratings on share white. Test score levels and progress are estimated as described in Appendix B.1. The best linear predictor is the fitted value obtained from model (6) augmented with a sector dummy. Race-balanced progress is the residual from a regression of progress on share white. Robust standard errors reported in parentheses.

Appendix Figures and Tables

Figure A1. Adjusted Ratings and Race



New York

B. Denver

Notes: These binned scatterplots depict the relationship between three sorts of progress ratings and the share of students at a school that are white. Red triangles correspond to the benchmark progress rating, while green squares correspond to the racially-balanced progress rating obtained as the residual from equation (3). Orange diamonds correspond to the best linear predictor of school value-added, obtained as the fitted values from (6) augmented with a sector dummy. Bins are defined by 0.1 increments in share White with the last bin grouping schools with share white ≥ 0.6. Ratings are mean zero and scaled to have standard deviation equal to the standard deviation of school quality across schools in the district.

Table A1. Descriptive Statistics

	New York		Denver	
	All	With risk	All	With risk
	(1)	(2)	(3)	(4)
Demographics				
Hispanic	0.413	0.445	0.592	0.581
Black	0.231	0.254	0.125	0.140
Asian	0.184	0.171	0.032	0.033
White	0.154	0.110	0.210	0.201
Female	0.494	0.484	0.493	0.494
Free/reduced price lunch	0.731	0.763	0.723	0.703
Special education	0.201	0.215	0.102	0.087
English language learner	0.113	0.113	0.393	0.416
Baseline scores				
Math (standardized)	0.000	-0.063	0.000	0.077
ELA (standardized)	0.000	-0.055	0.000	0.070
Enrollment				
Screened	0.067	0.044	0.000	0.000
Lottery	0.933	0.956	1.000	1.000
Share non-compliant	0.268	0.324	0.300	0.291
Share not offered	0.149	0.134	0.182	0.048
Students	184,760	46,095	37,089	8,100
Schools	624	594	80	75
Lotteries (schools with risk)		448		67

Notes: This table describes the Denver and New York student samples used to compute ratings and estimate school quality. Column 1 show statistics for New York middle school students enrolled in 6th grade in the 2016-17 through 2018-19 school years. Column 3 shows descriptive statistics for Denver students enrolled in 6th grade in the 2012-13 through 2018-19 school years. Columns 2 and 4 describe the corresponding samples of applicants with assignment risk at at least one school. Baseline characteristics and lagged scores are from 5th grade. Baseline scores are standardized to be mean zero and standard deviation one in the student-level test score distribution, separately by year. Screened schools are defined as schools without any lottery programs. The share non-compliant is defined as the proportion of students who enroll other than where offered a seat; this includes students receiving no offers.

Table A2. School Counts

	New York				Denver		
	TPS						
	Non-screened	Screened	Charter	All schools	TPS	Charter	All schools
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. School-year counts							
In sample	1359	142		1501	223	150	373
Not in sample	80	3		83	52	10	62
Total	1439	145		1584	275	160	435
Panel B. School counts (2016)							
In sample	433	47	90	570	31	22	53
Not in sample	17	0	28	45	9	2	11
Total	450	47	118	615	40	24	64

Notes: This table describes the schools in the IV estimation sample. These schools enroll at least one student with non-degenerate risk. The columns labelled "TPS" indicate traditional public schools. Screened schools in New York are schools that offer only screened programs. In New York, student-level charter enrollment is only observed in the 2016-2017 school year. In Panel A, charter school-years are counted as non-screened observations.

Table A3. Tests for Differential Attrition

	New York (1)	Denver (2)
Offered progress	0.032 (0.019)	0.022 (0.038)
N	53,094	9,234
Mean follow-up rate	0.898	0.896

Notes: This table reports differential attrition estimates. These estimates come from regressions of a follow-up indicator on the estimated progress rating of the offered school, controlling for expected progress rating and running variable controls in the New York sample. Robust standard errors are reported in parentheses.

Table A4. Projections of School Quality and School Ratings on Share White and Asian

		Test score levels			Test score progress		
	Value-added projection (derived)	Value-added projection (derived)	Rating projection (OLS)	Value-added projection (derived)	Value-added projection (derived)	Rating projection (OLS)	Value-added projection (derived)
Dependent variable:	School quality (β)	School quality (β)	Test score levels (R)	School quality (β)	School quality (β)	Test score progress (R)	School quality (β)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Panel A. New York				
Test score levels		0.164 (0.055)		0.536 (0.071)			
Test score progress					0.738 (0.037)		0.812 (0.038)
Screened school dummy	-0.047 (0.035)		0.101 (0.012)	-0.101 (0.035)		-0.037 (0.016)	-0.017 (0.032)
Share white and Asian	-0.046 (0.046)		0.541 (0.013)	-0.336 (0.062)		0.199 (0.016)	-0.207 (0.045)
First-stage F				15.1			
N (school-year)				1501			
			Panel B. Denver				
Test score levels		0.482 (0.148)		1.37 (0.221)			
Test score progress					0.843 (0.089)		0.945 (0.097)
Charter school dummy	0.100 (0.036)		0.099 (0.011)	-0.033 (0.045)		0.139 (0.020)	-0.033 (0.038)
Share white and Asian	0.175 (0.126)		0.834 (0.025)	-0.977 (0.219)		0.405 (0.049)	-0.210 (0.122)
First-stage F				9.09			
N (school-year)				373			

Notes: This table reports estimates from projections of levels and progress school ratings and causal value added on school characteristics, including the share white and Asian. The models and derivation procedure used to compute these estimates are as the estimates in Table 2. Robust standard errors are reported in parentheses.

Table A5. Projections of School Quality and School Quality on Share Non-FRPL

		Test score levels			Test score progress		
	Value-added projection (derived)	Value-added projection (derived)	Rating projection (OLS)	Value-added projection (derived)	Value-added projection (derived)	Rating projection (OLS)	Value-added projection (derived)
Dependent variable:	School quality (β)	School quality (β)	Test score levels (R)	School quality (β)	School quality (β)	Test score progress (R)	School quality (β)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Panel A. New York				
Test score levels		0.232		0.451			
		(0.052)		(0.063)			
Test score progress					0.761		0.774
					(0.037)		(0.037)
Screened school dummy	-0.050		0.060	-0.077		-0.040	-0.019
	(0.035)		(0.013)	(0.035)		(0.016)	(0.032)
Share non-FRPL	0.018		0.656	-0.278		0.144	-0.094
	(0.050)		(0.018)	(0.059)		(0.024)	(0.046)
First-stage F				20.4			
N (school-year)				1501			
			Panel B. Denver				
Test score levels		0.443		1.29			
		(0.147)		(0.213)			
Test score progress					0.851		0.941
					(0.083)		(0.096)
Charter school dummy	0.087		0.066	-0.011		0.124	-0.018
	(0.036)		(0.012)	(0.041)		(0.020)	(0.037)
Share non-FRPL	0.151		0.745	-0.842		0.344	-0.178
	(0.112)		(0.023)	(0.188)		(0.044)	(0.109)
First-stage F				10.9			
N (school-year)				373			

Notes: This table reports estimates from projections of levels and progress school ratings and causal value added on school characteristics, including the share not eligible for a free or reduced-price lunch. The models and derivation procedure used to compute these estimates are as the estimates in Table 2. Robust standard errors are reported in parentheses.

Table A6. IV VAM Regressions

	Over-identified (school assignment instruments)		Just-identified (offered mediator instruments)	
	NYC	Denver	NYC	Denver
	(1)	(2)	(3)	(4)
Mediators				
Test score levels	-0.140	0.417	-0.234	-0.006
	(0.064)	(0.230)	(0.102)	(0.437)
Test score progress	0.839	0.847	1.10	1.05
	(0.044)	(0.116)	(0.064)	(0.151)
Screened school dummy	-0.009		0.011	
	(0.033)		(0.037)	
Charter school dummy		-0.066		0.010
		(0.044)		(0.063)
Share white	-0.087	-0.547	-0.051	-0.129
	(0.064)	(0.217)	(0.079)	(0.340)
First-stage F	23.2	15.1	608	31.7
Value-added std. dev.	0.194	0.217		
N	46,095	8,100	46,095	8,100

Notes: This table reports IV VAM parameter estimates. These estimates are used to obtain the estimates reported in Table 2. The set of listed mediators is treated as endogenous. Columns 1 and 2 use individual school assignment offer dummies as instruments for 2SLS estimation. Columns 3 and 4 use values of the mediator at the offered school as instruments. All models control for school assignment risk and year fixed effects fully interacted with the demographic variables listed in Appendix Table A1 and cubic functions of 5th grade math and ELA scores. New York models also include local linear functions of the relevant screened-school tie-breakers. Ratings are demeaned and scaled to have variance matching that of school quality across schools in the district. Robust standard errors are reported in parentheses.

Table A7. Tests for Equality of IV and OLS Estimates of Racial Imbalance

	Hausman		Joint estimation	
	Test score levels	Test score progress	Test score levels	Test score progress
	(1)	(2)	(3)	(4)
	Panel A: New York			
Racial imbalance				
IV (school quality)	0.004		0.004	
	(0.061)		(0.062)	
OLS	0.687	0.222	0.687	0.222
	(0.024)	(0.026)	(0.024)	(0.026)
IV - OLS	-0.683	-0.219	-0.683	-0.219
	(0.055)	(0.055)	(0.066)	(0.068)
	[0.000]	[0.000]	[0.000]	[0.001]
	Panel B: Denver			
Racial imbalance				
IV (school quality)	0.188		0.188	
	(0.135)		(0.122)	
OLS	0.881	0.433	0.881	0.433
	(0.027)	(0.051)	(0.027)	(0.051)
IV - OLS	-0.693	-0.246	-0.693	-0.246
	(0.132)	(0.125)	(0.125)	(0.131)
	[0.000]	[0.049]	[0.000]	[0.060]

Notes: This table reports tests for equality between the IV estimates of the racial imbalance of school quality and OLS estimates of the racial imbalance of either the levels rating or the progress rating. Columns 1 and 2 use a Hausman (1978) test which takes as the covariance between the IV and OLS estimators the variance of the OLS estimator. Columns 3 and 4 compute the covariance between the IV and OLS estimators by jointly estimating these models. Standard errors, clustered by school-year, are reported in parentheses. P-values for the test of IV and OLS equality are reported in brackets.

Table A8. Comparison of Racial Imbalance in GreatSchools' Levels and Progress Ratings

	Test score levels (1)	Test score progress (2)
Panel A: USA		
Charter school dummy	0.019	0.015
	(0.005)	(0.006)
Share white	0.632	0.310
	(0.004)	(0.006)
N (schools)	72573	61247
Panel B: New York State		
Charter school dummy	-	-
Share white	0.625	0.095
	(0.022)	(0.030)
N (schools)	3979	3099
Panel C: Colorado		
Charter school dummy	0.019	0.015
	(0.005)	(0.006)
Share white	0.735	0.302
	(0.022)	(0.031)
N (schools)	1210	1474

Notes: This table reports racial imbalance regressions for GreatSchools levels and progress ratings in the 2018 school year. Panel A includes all public schools in the United States with GreatSchools ratings, while Panels B and C restrict the sample to schools in New York state and Colorado, respectively. Ratings are standardized by state to have mean zero and standard deviation 0.2, roughly the standard deviation of school quality in both NYC and Denver. All models include district fixed effects, which absorb charter school indicators in New York. Levels is GreatSchools' Test Score Rating, and progress is GreatSchools' Student Progress Rating when available and Academic Progress Rating otherwise. See Appendix B.1 and `https://www.greatschools.org/gk/ratings-methodology/` for more information on GreatSchools ratings.

Table A9. Centralized Assignment in Large Public School Districts

	All (1)	Minority (2)
All districts		
Enrollment (% of all districts)	100%	91%
N	100	87
Centralized		
Enrollment (% of all districts)	36%	34%
N	26	24
Partially centralized		
Enrollment (% of all districts)	69%	65%
N	59	52
Any randomness		
Enrollment (% of all districts)	83%	77%
N	76	66

Notes: This table describes the student assignment mechanism for the 100 largest public school districts in the United States. Column 2 considers districts enrolling at least 30% Black and Hispanic students. Centralized districts employ mechanisms with quasi-random offer variation for traditional public schools. Partially centralized districts include those with a centralized aftermarket for school choice transfers away from neighborhood schools. Any randomness districts employ mechanisms with any random offer variation, for instance decentralized lotteries at non-traditional public schools. Further details on definitions and coding procedures are available on request. Enrollment data reflect fall 2019 figures from the NCES.

B Data Appendix

B.1 School Quality Measures

The measures used here are motivated by the "test score" and "progress" ratings published by GreatSchools.org. The test score rating is a levels measure that uses student proficiency rates as inputs. The progress rating uses state-reported estimates of student growth as inputs. Our progress ratings are based on models underlying the "growth" rating reported by Colorado and the student growth percentile estimates reported by New York.[23]

Our computation differs in a few ways from GreatSchools and state ratings because we are interested in sixth-grade ratings for specific years and outcomes; it's sometimes unclear which grades and years were used to compute published ratings. Also, GreatSchools ratings transform state-reported inputs into a discrete 1-10 rating; we omit this step. Like GreatSchools ratings, our computation is year-specific.[24]

Our *levels* rating averages the share of students who are proficient in math and the share of students who are proficient in English Language Arts (ELA), as measured by sixth-grade achievement tests. Formally, this is $R_j = (E[q_i^m \mid D_{ij} = 1] + E[q_i^e \mid D_{ij} = 1])/2$, where q_i^s indicates a student who is deemed proficient in subject s (math or ELA). Students are deemed proficient when their scores cross state-determined cutoffs.

Our *progress* rating is derived from estimates of student growth percentile models. Neither of these procedures involve simple difference-based measures of growth; rather, they adjust for lagged achievement. Nevertheless, the resulting measures are often called a "student growth percentile," or SGP (Castellano and Ho, 2013). The underlying models are described in New York State Education Department (2020) for New York and Colorado Department of Education (2019) for Colorado.

For purposes of our analysis, New York growth percentiles are computed by first estimating the regression:

$$Y_i^s = \delta^s + X_i'\Gamma^s + \eta_i^s,$$

for each subject $s \in \{m, e\}$. Here X_i is a control vector including 3rd, 4th, and 5th grade achievement scores. Missing lagged scores are coded to zero, with indicators for missing scores also included in X_i. From these regressions we compute the percentile rank, r_i^s, of the residual η_i^s in the city distribution of students. The progress rating is then the mean of the

[23] These ratings can be found through Colorado's Performance Snapshot (`https://www.cde.state.co.us/code/accountability-performancesnapshot`) and the "ACC EM Growth" table in New York's Report Card Database (`https://data.nysed.gov/downloads.php`).

[24] See `https://www.greatschools.org/gk/ratings-methodology/` for more information on the GreatSchools ratings computation.

school average math and ELA ranks: $R_j = (E[r_i^m \mid D_{ij} = 1] + E[r_i^e \mid D_{ij} = 1])/2$.

Student growth percentiles for Denver are computed using quantile regression. This procedure begins by using quantile regression to fit conditional quantiles as a function of the control vector, X_i, listed above. Quantile regression coefficients are computed for every percentile from 1-99. The Denver percentile rank is the quantile value, τ, that minimizes $Y_i^s - X_i'\hat{\Gamma}_\tau^s$, where $\hat{\Gamma}_\tau^s$ is the estimated vector of quantile regression coefficients for percentile τ. As in New York, subject-specific results are averaged to produce a single progress rating for each school and year.

B.2 Standardization of Outcomes and Ratings

The primary outcome for our analysis is constructed by first summing each student's scaled math and ELA sixth-grade test scores, then standardizing this sum to have mean zero and standard deviation one, separately by city and year. Year-specific school value added, β_j, is therefore measured in units of student-level test score standard deviations.

To facilitate comparisons of forecast coefficients across ratings, alternative ratings are scaled to have the same standard deviation as causal value-added. Specifically, we estimate the IV VAM model (9) and use the results to form an estimate $\hat{\sigma}_\beta$ of the standard deviation of causal value-added, as described in Angrist et al. (2021). For each year, we then multiply each rating (deviated from its mean) by the ratio of $\hat{\sigma}_\beta$ to its own standard deviation. This results in a rating with mean zero and standard deviation $\hat{\sigma}_\beta$. The forecast coefficients in Table 2 can therefore be interpreted as gains in standard deviations of causal value-added associated with a one standard-deviation increase in school ratings. A rating that accurately orders schools according to causal value-added should be expected to generate a forecast coefficient of roughly unity. It's worth noting, however, that the forecast coefficient may not be exactly one even for a rating that ranks schools exactly in order of β_j, since value-added and school ratings are measured in different units, even after rescaling.