Blueprint Labs

Discussion Paper #2022.01

Race and the Mismeasure of School Quality

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

January 2022

MIT Department of Economics
77 Massachusetts Avenue, Bldg. E53-390
Cambridge, MA 02139

National Bureau of Economic Research
1050 Massachusetts Avenue, 3rd Floor
Cambridge, MA 02138

Race and the Mismeasure of School Quality[*]

Joshua Angrist[†] Peter Hull[‡] Parag A. Pathak[§]

Christopher R. Walters[¶]

January 14, 2022

Abstract

In large urban districts, schools enrolling more white students tend to have higher school performance ratings. We use an instrumental variables strategy leveraging centralized school assignment to identify the drivers of the correlation between racial make-up and ratings. Estimates from Denver and New York City suggest the relationship between widely-reported 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 causal value-added as well or better than the corresponding unadjusted measures.

[*] 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 go to Eryn Heying, Anna Vallee, and Talia Gerstle for invaluable administrative support. Jimmy Chin, Ray Han, and Ini Umosen provided expert research assistance. The research described here was carried out under data-use agreements between MIT and the New York City and Denver Public School districts.

[†] MIT and NBER. Email: angrist@mit.edu

[‡] Brown University and NBER. Email: peter_hull@brown.edu

[§] MIT and NBER. Email: ppathak@mit.edu

[¶] University of California, Berkeley and NBER. Email: 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 provided by GreatSchools.org and various state accountability offices, meet the demand for information on school quality from both parents and policymakers. The intense public interest in school performance is manifest in the fact that real estate sites like Zillow and Redfin feature school ratings prominently. School performance ratings appear to affect families' choices of where to live and where to enroll (Bergman and Hill, 2018; Hasan and Kumar, 2019), as well as district decisions related to school closures, takeovers, and expansions (Rockoff and Turner, 2010; Abdulkadiroğlu et al., 2016; Cohodes et al., 2021).

Do highly sought-after school ratings serve the public interest? Barnum and LeMee (2019) and other journalists note 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, the student body enrolled 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 this suggest that links between published school ratings and racial composition may contribute to ongoing 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 first brought to economists' attention by Welch (1973). Improvements in the quality of predominately-Black schools account for much of the reduction in Black-white wage gaps seen in the 1950s through the 1970s (Card and Krueger, 1992a,b). This progress notwithstanding, school attendance remains highly segregated, even within school districts (Monarrez, 2021). The higher achievement and graduation rates found at schools that enroll more white students may reflect these schools' greater impact on learning. 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 reflect this view (Jonas, 2021).

While the link between school rankings and schools' racial make-up may reflect differences in quality, this relationship 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

[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.

of the schools they attend. School ratings that fail to adjust for this differential conflate differences in school quality with differences in student composition; recent research suggests such selection bias is pervasive (Angrist et al., 2017; Abdulkadiroğlu et al., 2020b). Rating schemes that reward family background rather than educational effectiveness are likely to direct households to low-minority rather than higher-quality schools, while penalizing schools that improve achievement for less-advantaged groups.[2]

This paper investigates the relationship between widely-used public school ratings and student racial composition, drawing broader implications for school assessment systems. Our analysis focuses on two properties of a school rating: *predictive accuracy*, defined as its (squared) correlation with a school's causal effect on achievement, and *racial imbalance*, defined by the 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 appear to face an unavoidable trade-off between predicative accuracy and racial imbalance.

Our findings show that the trade-off between predictive accuracy and racial imbalance is much smaller than the observed correlation between school ratings and racial composition suggests. This conclusion is reached in two steps. First, we present a simple but novel characterization of the potential link between predictive accuracy and racial imbalance. School quality is not directly observed, so this trade-off formula is not immediately applicable. As in Abdulkadiroğlu et al. (2017, 2020a), we surmount this identification challenge by using the random variation in school attendance generated by centralized school assignment systems. Building on this framework and the instrumental variables value-added model (IV VAM) approach from Angrist et al. (2021), we derive feasible estimators of the relationships between causal value-added, racial composition, and conventional school ratings.

The IV VAM estimates are used to quantify the trade-off between predictive accuracy and racial imbalance for middle school students in New York City (NYC) and Denver. Both districts allocate seats using a centralized match that generates partially randomized variation in school assignment, yielding the instruments needed for IV VAM. These two districts are also central to discussions of segregation and school access. New York is important because it's one of America's largest districts and because it has a long history of de facto segregation. Denver draws attention because it's both a majority Hispanic district and a pioneer of unified school enrollment, with one centralized match allocating seats at all publicly funded schools (including charter schools).

School performance ratings based on achievement levels and on achievement growth are

[2] Compounding this concern, Hart and Figlio (2015) find that the enrollment decisions of highly educated parents respond more to school quality ratings than do decisions of less-educated parents.

both highly correlated with schools' racial composition in NYC and Denver. Our analysis substantiates the view that this correlation is largely an artifact of selection bias. Specifically, IV VAM estimates reveal that causal value-added is statistically unrelated to racial composition in both cities. The weak relationship between school value-added and racial composition suggests, given our theoretical characterization, that adjusting school ratings to reduce racial imbalance may come at little cost. We confirm this prediction by showing that a conventional progress-based rating adjusted to be uncorrelated with race has predictive accuracy no worse than (and sometimes better than) that of the corresponding unadjusted measures. Moreover, in both NYC 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.

The rest of this paper is organized as follows. Section 2 describes our school district settings and data. Section 3 develops the IV VAM econometric framework as applied to the trade-off between predictive accuracy and racial imbalance. Section 4 presents our findings and Section 5 concludes with directions for further work.

2 Settings and Data

Our Denver analysis sample includes students applying for sixth-grade seats at any of the Denver Public School (DPS) district's middle schools between the 2012-2013 and 2018-2019 school years. Our NYC analysis sample includes sixth-grade applicants to NYC middle schools for the 2016-2017 through 2018-2019 school years. We observe the school preferences and priorities submitted by each applicant and the subsequent assignments generated by each district's centralized school assignment system. We also have data on subsequent school enrollment, student demographics, and achievement scores.[3] Denver outcomes are from the Colorado Student Assessment Program (CSAP) and Colorado Measures of Academic Success (CMAS) standardized tests. NYC outcomes come from New York state achievement tests. The main dependent variable used in our analysis is the sum of a student's scaled math and ELA scores in sixth grade. We standardize this sum to have mean zero and standard deviation one in each city, separately by year. Combining math and ELA scores helps to align our outcome with ratings reported by GreatSchools.org, school districts, and states, which also use both subjects.

Students in Denver rank up to five schools among those participating in the DPS unified enrollment match. Priorities are assigned based on criteria like sibling status and the applicant's residential neighborhood. A deferred acceptance (DA) algorithm implemented with a

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

single lottery tie-breaker assigns students to schools. NYC middle school applicants rank up to 12 academic programs; schools may host more than one program. For the purposes of our analysis, multiple programs are aggregated to the school level. The NYC match 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.

As in the Angrist et al. (2021) study of school value-added, our empirical strategy leverages the randomness embedded in each city's school assignment mechanism. We follow Abdulkadiroğlu et al. (2017, 2020a) 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 analytical large-market approximation to the school assignment propensity score for DA with a lottery tie-breaker.[4] Assignment risk for NYC applicants is computed as described in Abdulkadiroğlu et al. (2020a). NYC assignment risk depends in part on bandwidths for screened school tie-breakers, similar to those used in standard regression discontinuity designs.[5] Score conditioning yields a stratified randomized trial: conditional on assignment risk, school assignment is independent of applicant characteristics, both observed and unobserved (this is an application of the Rosenbaum and Rubin (1983) propensity score theorem).

Our analysis of school ratings focuses on two achievement-based measures of school quality meant to replicate widely-disseminated state ratings for Colorado and New York state. The *levels* rating used here consists of the share of students scored as proficient on state assessments, averaged across math and English language arts (ELA) tests. The *progress* rating used here is 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 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 estimate of the distribution of school value-added, detailed below. Appendix A.1 describes the construction of these school ratings along with the procedures used to standardize outcomes and ratings.

Appendix Table A1 describes the students and schools in the DPS and NYC samples, separately for the full sample of enrolled DPS or NYC middle school students and for school

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

[5] The NYC score is the *local DA score* described in Section 4.2 of Abdulkadiroğlu et al. (2020a). Bandwidths used here are computed as suggested by Calonico et al. (2019).

match applicants with non-degenerate assignment risk. This sample of applicants, indexed by i, have a propensity score p_{ij} strictly between zero and one for at least one school j. As is typical of large urban districts, most DPS and NYC students are from disadvantaged backgrounds, with over 70 percent eligible for a subsidized lunch. Roughly a quarter of the students in each sample face some assignment risk. 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.

Appendix Table A2 validates the natural experiment generated by centralized assignment by comparing the characteristics of students offered seats at higher-rated and lower-rated schools (these comparisons are based on the progress rating). Uncontrolled comparisons show large differences in student characteristics between those offered seats at high- and low-rated schools, but these differences vanish when we adjust for assignment risk. The fact that risk adjustment balances observed characteristics by offer status suggests unobserved characteristics are likely balanced as well.[6]

Figure 1 shows that both the levels and progress ratings are highly correlated with the racial composition of schools. 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 achievement ratings. In NYC, a regression of levels ratings on share white yields a slope coefficient of 0.64 with a robust standard error of 0.02. 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 0.3 standard deviations. The corresponding slope is smaller for progress, falling to 0.20, but the relationship remains clear and statistically precise. Evidence of racial imbalance for Denver is remarkably similar, with coefficients of 0.76 for the levels rating and 0.29 for the progress rating (both again 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

[6] 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 (2020). Appendix Table A3 further shows that differential attrition is unlikely a concern in this sample: follow-up rates for key outcomes are largely unrelated to assigned school ratings, conditional on assignment risk.

school $j \in \{1, ..., J\}$, denoted Y_{ij}, is given as:

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

where parameter $\beta_j \equiv E[Y_{ij}]$ is the contribution of attendance at school j to achievement. We refer to β_j as school j's *quality* or *value-added*, reserving these terms for parameters determining causal school effects. 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 the analysis on selection bias rather than treatment-effect heterogeneity.

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

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

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 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 distinguished by the demographic composition of their student bodies as well as by their value-added. Let W_j denote the share of students enrolled in school j designated as white (or any other race). Specifically, $W_j = E[w_i \mid D_{ij} = 1]$, 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 school rating accurately reveals a demographic gap in school quality. Alternatively, this correlation may arise 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 considers the following 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 of a rating scheme are defined for any choice of R_j, so that $\mathcal{I}_\beta$ denotes the slope coefficient from a regression of β_j on W_j.[7]

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:

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

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

Although racial imbalance is easily eliminated, balance may come at the cost of reduced predictive accuracy. To describe this trade-off, consider first the coefficients on ratings in the following two regression models 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, $\tilde{R}_j$. As in Angrist et al. (2021), both φ and $\tilde{\varphi}$ are *forecast coefficients*, quantifying the relationship between school quality and imperfect ratings.

Suppose that schools with a higher share of white students tend to be ranked higher, as in Figure 1. The two forecast coefficients defined above 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 $Corr(R_j, W_j) > 0$. □

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. Negative τ corresponds to a scenario in which schools with a

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

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 adjustment on predictive accuracy is given 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 $\mathcal{I}_R > 0$. Then $\rho_{\tilde{R}} > \rho_R$ if and only if $\mathcal{I}_\beta < -\tau(\varphi/\tilde{\varphi})$.*

Proof. See Appendix A.2. □

This result is especially sharp in a scenario where school quality is unrelated to race, so $\mathcal{I}_\beta = 0$. In this case, as long as ratings are racially imbalanced ($\mathcal{I}_R > 0$) but still informative, then $\tau < 0$ and we are sure that $\rho_{\tilde{R}} > \rho_R$.[8] 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 solely interested in maximizing predictive accuracy might combine information on racial make-up with ratings data, with a rating given by the fitted value 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 R_j and $\tilde{R}_j$ since the extra regressor, W_j, cannot reduce r-squared.[9] 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 through 3 show that the trade-off between the predictive power and racial imbalance of a school rating scheme depends on the two forecast coefficients

[8] 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 slope coefficient from a regression of W_j on R_j. When $\mathcal{I}_R > 0$, $\alpha > 0$, so $\tau < 0$ when $\varphi \propto Cov(\beta_j, R_j) > 0$.

[9] To see this for $\tilde{R}_j$, write (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 $\tilde{R}_j$.

defined above, the sign of τ in equation (5), and the racial imbalance of value-added. The challenge in applying these results is that the set of school quality parameters, β_j, are unobserved. This challenge notwithstanding, we can estimate the determinants of predictive accuracy and racial imbalance for alternative ratings using the IV VAM empirical strategy introduced in Angrist et al. (2021). Our application of IV VAM uses instrumental variables to estimate the coefficients in (4) and (5), that is, φ, $\tilde{\varphi}$, and τ. IV VAM also yields a measure of $\mathcal{I}_\beta$, the slope from a regression of school quality on share white, as well as an estimate of the total variance of β_j, which is used to calculate the predictive accuracy of each rating.

3.2 Identification and Estimation

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

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

As detailed below, the vector M_j includes a constant, school ratings, share white, school sector dummies, and outside estimates of value-added. Forecast regression (7) is a linear projection, so $E[M_j\xi_j] = 0$ by definition of the 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 regression model, but equation (8) need not be: selection bias makes it likely that 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 $M_{j(i)}$.

The IV VAM estimating equation includes a vector of individual-level control variables, X_i, including school assignment risk and other applicant characteristics. Control for the latter isn't necessary for identification, but enhances precision.[10] 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

[10] 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 NYC include local linear functions of the relevant screened-school tie-breakers; see Abdulkadiroğlu et al. (2020a) for details.

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 applicant heterogeneity, ε_i. The former reflects determinants of value-added not explained by the included endogenous variables, 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. Intuitively, these conditions require the relationships between individual school offers and residual school quality to average to zero over schools.

IV exclusion restrictions are made more plausible by including strong predictors of school quality in M_j. Intuitively, adding such mediators reduces and perhaps even eliminates variation in residual school quality, ξ_j. In our implementation, M_j includes both the levels and progress ratings, share white, a dummy for charter schools (in Denver), a dummy for screened schools (in NYC), and risk-controlled value-added (RC VAM) school quality estimates. As detailed in Angrist et al. (2021), RC VAM uses ordinary least squares (OLS) to estimate value-added in a regression model with controls for demographic characteristics, lagged test scores, and assignment risk. RC VAM appears to predict school quality in Denver and New York well, supporting IV VAM exclusion restrictions.

The parameters in (9) are estimated by two-stage least squares (2SLS), separately for each city. 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 the mediators 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 M_{2j} collecting the other elements of M_j and $\psi = (\psi'_1, \psi'_2)'$ partitioned conformably. 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, 2SLS estimation of (9) also generates an estimate of the variance of β_j, needed to compute predictive accuracy. (In practice, the variance of β_j is well-approximated by the variance of the estimated $M'_j\psi$.)

4 Results

School quality is unrelated to share white in the sample of New York City middle schools. This can be seen in the first column of Panel A in Table 1, which reports estimates of the projection of β_j on share white and a screened school indicator for schools in NYC. The full set of IV VAM estimates underlying these results appears in Appendix Table A6.[11] Both of these coefficient estimates are small and neither is significantly different from zero. The estimates reported in column 3, by contrast, show that share white and screened school status are highly predictive of school ratings based on test score levels—a pattern consistent with Figure 1. Together, the results in columns 1 and 3 imply that the strong relationship between school ratings and share white in NYC reflects selection bias.[12]

Levels ratings are weakly related to school quality in NYC. Specifically, the estimated forecast coefficient in column 2 of Table 1 shows that a one standard deviation improvement in test score levels is associated with only a 0.21 standard deviation increase in causal value-added.[13] 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 conforms to the pattern discussed in the previous section: schools that enroll more white students, as well as highly sought-after screened schools, are of lower quality than other similarly-rated schools.

As can be seen in column 5 of Table 1, progress ratings predict NYC school quality with a forecast coefficient of about 0.77, 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 0.22 coefficient for share white in a regression of progress on share white and a screened-school dummy. As can be seen in column 7, when quality is predicted by progress and share white, the progress coefficient remains high, 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 unrelated, so that disproportionately white and screened schools are, on

[11] The first-stage F-statistics for these estimates are close to the rule-of-thumb threshold of 10 commonly used to diagnose weak instruments. But the 2SLS estimates in the table are close to estimates from a just-identified IV estimator, displayed nearby, which yields considerably higher first-stage F statistics. The just-identified estimator replaces individual school offer dummies in the instrument list with values of the mediator at the offered school as instruments, one for each mediator.

[12] Appendix Tables A4 and A5 report results from models that replace share white with share white or Asian and with the share of students eligible for a free or reduced price lunch. These variations yield results similar to those in Table 1.

[13] As detailed in Appendix A.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.

average, over-rated.

Analogous results for Denver, reported in Panel B of Table 1, are qualitatively similar to those for New York, though these smaller-district estimates are less precise. Column 1 shows a modest positive but insignificant relationship between school quality and share white, while Denver's many charter schools generate a precisely estimated achievement gain of about 0.12 standard deviations. As in NYC, race predicts levels more than progress (compare columns 3 and 6 in Panel B), but both predictive relationships for ratings are strong. Also as in NYC, multivariate school quality projections for Denver yield negative (though relatively imprecise) estimated coefficients on share white when ratings are included as an explanatory variable; these results are reported in columns 4 and 7 of Panel B.

Figure 2 highlights implications of the results in Table 1 by plotting alternative ratings against share white in both NYC and Denver schools. 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 seen previously in Figure 1, the relationship between the progress rating and share white for NYC schools is clearly upward-sloping (the y-axis range in Figure 2 is half that in the first figure). 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 NYC school quality given the progress rating, share white, and screened school status (the fitted value from the model generating column 7 of Table 1) yields a CEF close to that for race-balanced ratings.

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 NYC (compare the estimates for the two cities in column 6 of Table 1). Consistent with these estimates, the CEF for the best linear predictor of Denver school quality plotted in Panel B of Figure 2 is weakly dependent on share white. Even so, the best linear predictor for Denver school quality rises much less steeply in share white than does the CEF of the raw Denver progress rating.

Table 2 summarizes this investigation with estimates of predictive accuracy (ρ_R) and racial imbalance ($\mathcal{I}_R$) for alternative ratings. In both NYC and Denver, progress ratings are far more accurate than levels ratings. Progress ratings are also much more weakly correlated with share white than levels ratings. This improvement notwithstanding, progress remains substantially correlated with race. Race-balanced ratings boost predictive accuracy of NYC school quality, while leaving predictive accuracy for Denver schools virtually unchanged. The best linear predictor of school quality given progress ratings and share white 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 on race, if at all.

5 Conclusions

The oft-noted correlation between school ratings and racial composition raises the concern that such ratings promote segregation and penalize schools that serve minority students. At the same time, demographic differences in ratings may also signal important disparities in school quality. Our analysis uses the random assignment embedded in centralized assignment mechanisms to disentangle the relationship between school ratings, school quality, and race. We show that for middle schools in Denver and New York City, the fact that schools with more white students are highly rated reflects selection bias rather than educational quality. As a result, ratings purged of their correlation with share white predict school quality as well or better than standard measures based on achievement levels and progress.[14]

Denver and NYC share important features with other large urban districts, suggesting the patterns uncovered here may not be unique to these cities. We hope to explore the trade-off between predictive accuracy and racial imbalance elsewhere in the near future. Our analysis leaves open the question of how racially-balanced school ratings might affect household decision-making. Households appear to respond to information about school performance (Hastings and Weinstein, 2008; Bergman and Hill, 2018; Bergman et al., 2020; Houston and Henig, 2021; Campos and Kearns, 2021). Credible racially-balanced quality information may therefore increase the demand for 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., 2020b). A gauge of the relative importance of objective quality measures and peer characteristics in the school choice nexus remains a high priority for future work.

[14] 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`).

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.

——— (2020a): "Breaking Ties: Regression Discontinuity Design Meets Market Design," Cowles Foundation Working Paper 2170R. Forthcoming, *Econometrica.*

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020b): "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.

Borusyak, K. and P. Hull (2020): "Non-Random Exposure to Exogenous Shocks: Theory and Applications," NBER Working Paper No. 27845, September.

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 (2021): "The Impacts of Neighborhood School Choice: Evidence from Los Angeles' Zones of Choice," Working Paper.

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*.

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.

Hart, C. M. and D. N. Figlio (2015): "School Accountability and School Choice: Effects on Student Selection Across Schools," *National Tax Journal*, 68, 875–899.

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.

Houston, D. M. and J. R. Henig (2021): "The 'Good' Schools: Academic Performance Data, School Choice, and Segregation," Annenberg EdWorkingPaper No. 21-491.

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

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

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

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.

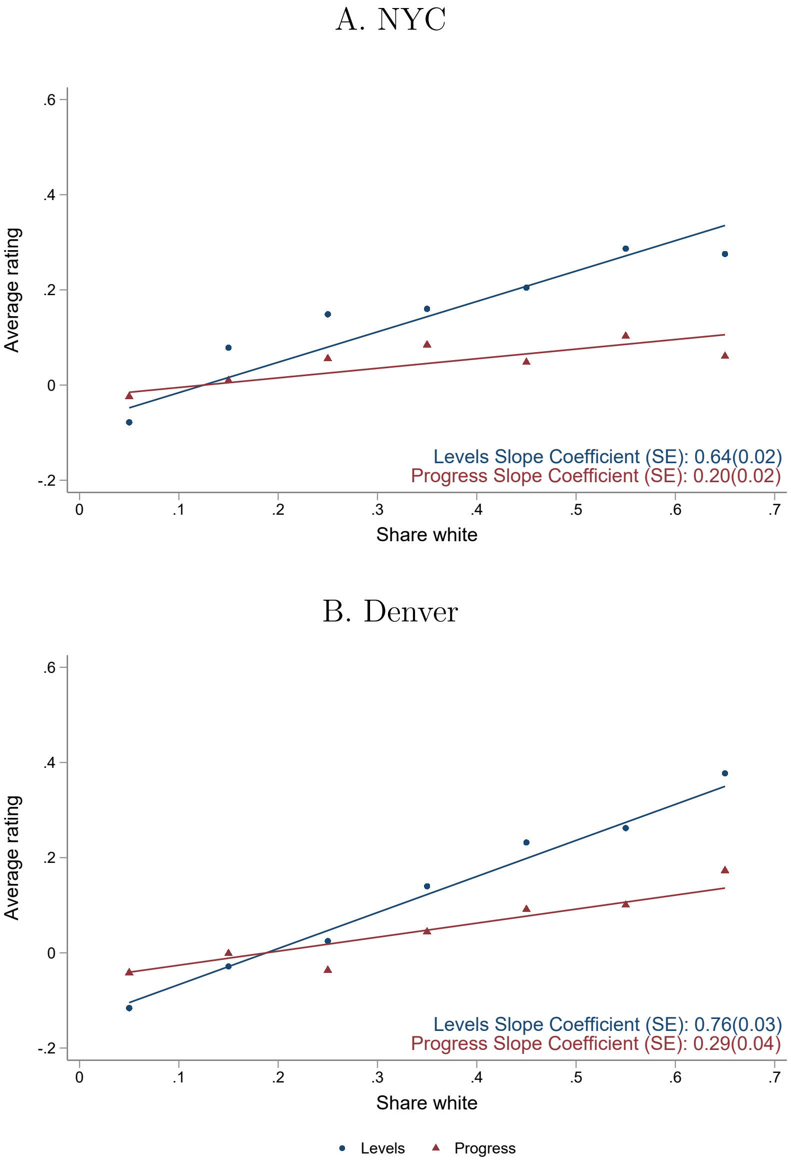
Rosenbaum, P. R. and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rothstein, J. 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

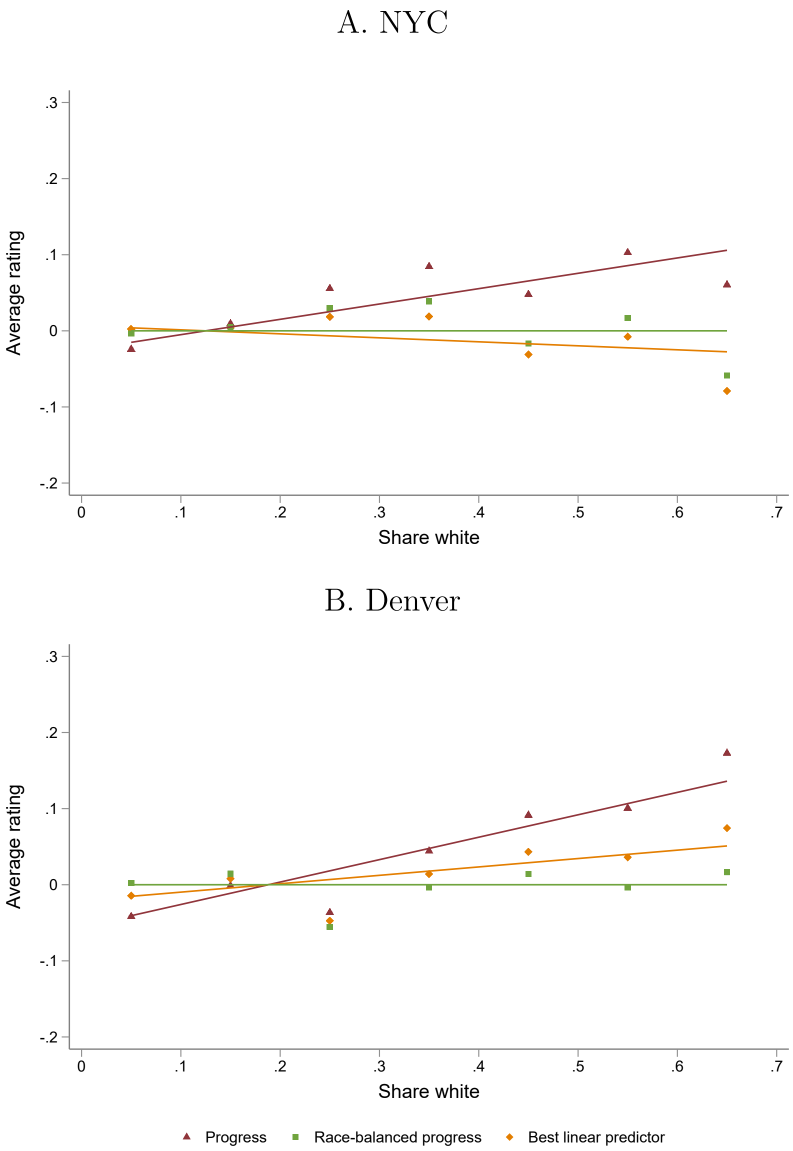


A. NYC

B. Denver

Notes: These 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 A.1. Ratings are mean-zero and scaled to have standard deviation equal to the standard deviation of school quality across schools in the district, which equals roughly 0.2 in both cities.

Figure 2. Adjusted Ratings and Race


A. NYC

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. 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.

Table 1. Projections of School Quality and School Ratings on School Characteristics

		Test score levels			Test score progress		
	VA projection (derived)	VA projection (derived)	Rating projection (OLS)	VA projection (derived)	VA projection (derived)	Rating projection (OLS)	VA 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. NYC			
Predictors							
Test score levels		0.208 (0.057)		0.417 (0.064)			
Test score progress					0.773 (0.041)		0.809 (0.041)
Screened school dummy	-0.052 (0.037)		0.100 (0.014)	-0.094 (0.037)		-0.034 (0.016)	-0.024 (0.034)
Share white	-0.034 (0.065)		0.683 (0.024)	-0.318 (0.073)		0.221 (0.025)	-0.213 (0.061)
N (school-year)				1501			
				Panel B. Denver			
Predictors							
Test score levels		0.506 (0.138)		1.29 (0.239)			
Test score progress					0.908 (0.096)		0.977 (0.113)
Charter school dummy	0.120 (0.037)		0.083 (0.010)	0.014 (0.046)		0.121 (0.018)	0.002 (0.041)
Share white	0.179 (0.137)		0.792 (0.028)	-0.841 (0.232)		0.364 (0.045)	-0.177 (0.136)
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 A6, 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 OLS regressions of school ratings on share white and a school sector dummy. Robust standard errors are reported in parentheses.

Table 2. Predictive Accuracy and Racial Imbalance

	NYC		Denver	
	Predictive accuracy (ρ)	Racial imbalance $(\mathcal{I})$	Predictive accuracy (ρ)	Racial imbalance $(\mathcal{I})$
	(1)	(2)	(3)	(4)
1. Test score levels	0.043	0.697	0.256	0.763
		(0.025)		(0.028)
2. Test score progress	0.597	0.216	0.825	0.323
		(0.026)		(0.045)
3. Race-balanced progress	0.632	0.000	0.835	0.000
		-		-
4. Best linear predictor	0.635	-0.041	0.859	0.138
		(0.065)		(0.136)

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

A Appendix

A.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.[15]

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.[16]

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 NYC and Colorado Department of Education (2019) for Colorado.

For purposes of our analysis, NYC 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

[15] 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`).

[16] 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 NYC, subject-specific results are averaged to produce a single progress rating for each school and year.

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 zero mean and standard deviation $\hat{\sigma}_\beta$. The forecast coefficients in Table 1 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.

A.2 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

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

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 $\lambda = \frac{Cov(W_j,R_j)}{Var(W_j)}$. With $Cov(W_j, R_j) > 0$, and hence $\lambda > 0$, equation (11) shows that $\rho_{\tilde{R}} > \rho_R$ if and only if

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

By the omitted variables bias formula $\tau + \tilde{\varphi}\lambda = \frac{Cov(W_j,\beta_j)}{Var(W_j)}$, completing the proof. □

A.3 Appendix Figures and Tables

Table A1. Descriptive Statistics

	NYC		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
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	46,099	37,089	8,100
Schools	624	594	80	75
Lotteries (schools with risk)		448		67

Notes: This table describes the Denver and NYC student samples used to compute ratings and estimate school quality. Column 1 show statistics for NYC 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. Statistical Tests for Balance

	NYC		Denver	
	Uncontrolled	Controlled	Uncontrolled	Controlled
	(1)	(2)	(3)	(4)
Demographics				
Hispanic				
Offered SGP	-0.202	0.030	-0.747	-0.043
	(0.008)	(0.028)	(0.020)	(0.070)
Any offer	-0.012	-0.011	-0.030	-0.009
	(0.003)	(0.009)	(0.007)	(0.029)
Black				
Offered SGP	-0.660	-0.004	0.040	0.059
	(0.007)	(0.025)	(0.014)	(0.048)
Any offer	-0.117	0.008	-0.040	-0.031
	(0.003)	(0.008)	(0.005)	(0.022)
White				
Offered SGP	0.440	-0.007	0.679	-0.065
	(0.006)	(0.016)	(0.019)	(0.061)
Any offer	0.061	0.005	0.083	0.072
	(0.002)	(0.005)	(0.005)	(0.019)
Female				
Offered SGP	0.017	0.024	-0.077	0.069
	(0.009)	(0.029)	(0.021)	(0.070)
Any offer	0.020	-0.020	0.009	-0.045
	(0.003)	(0.009)	(0.007)	(0.031)
Free/reduced price lunch				
Offered SGP	-0.333	0.037	-0.800	0.054
	(0.007)	(0.023)	(0.020)	(0.067)
Any offer	-0.077	-0.005	-0.074	-0.042
	(0.003)	(0.007)	(0.005)	(0.025)
Special education				
Offered SGP	-0.111	-0.010	-0.061	-0.012
	(0.007)	(0.023)	(0.012)	(0.038)
Any offer	-0.039	0.007	-0.001	0.031
	(0.003)	(0.008)	(0.004)	(0.019)
English language learner				
Offered SGP	0.021	0.023	-0.400	0.002
	(0.005)	(0.019)	(0.019)	(0.067)
Any offer	-0.016	0.001	-0.071	-0.087
	(0.002)	(0.006)	(0.007)	(0.031)
Baseline scores				
Math (standardized)				
Offered SGP	1.25	-0.005	1.36	0.175
	(0.016)	(0.052)	(0.042)	(0.142)
Any offer	0.273	-0.030	0.222	-0.028
	(0.006)	(0.017)	(0.012)	(0.057)
ELA (standardized)				
Offered SGP	0.939	-0.041	1.18	0.097
	(0.016)	(0.054)	(0.041)	(0.138)
Any offer	0.256	-0.009	0.190	-0.037
	(0.006)	(0.017)	(0.013)	(0.056)
N	184,760	46,099	37,089	8,100

Notes: This table reports balance statistics, estimated by regressing baseline covariates on the estimated student growth percentile of the offered school and an indicator for any offer. Columns 2 and 4 control for expected student growth percentile, any offer risk, and running variable controls in the NYC sample. Robust standard errors are reported in parentheses.

Table A3. Tests for Differential Attrition

	NYC	Denver
	(1)	(2)
Offered SGP	0.032	0.026
	(0.019)	(0.043)
Any offer	0.037	0.027
	(0.006)	(0.019)
N	53,098	9,234
Mean follow-up rate	0.898	0.896

Notes: This table reports differential attrition estimates. Estimates in column 1 are from regressions of a follow-up indicator on the estimated student growth percentile of the offered school, controlling for expected student growth percentile and running variable controls in the NYC 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		
	VA projection (derived)	VA projection (derived)	Rating projection (OLS)	VA projection (derived)	VA projection (derived)	Rating projection (OLS)	VA 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. NYC							
Predictors							
Test score levels		0.198		0.643			
		(0.060)		(0.078)			
Test score progress					0.774		0.856
					(0.041)		(0.042)
Screened school dummy	-0.050		0.097	-0.113		-0.038	-0.018
	(0.037)		(0.011)	(0.037)		(0.016)	(0.034)
Share white and Asian	-0.052		0.522	-0.387		0.197	-0.220
	(0.047)		(0.012)	(0.064)		(0.016)	(0.047)
N (school-year)				1501			
Panel B. Denver							
Predictors							
Test score levels		0.524		1.34			
		(0.137)		(0.261)			
Test score progress					0.906		0.968
					(0.098)		(0.110)
Charter school dummy	0.122		0.084	0.009		0.121	0.005
	(0.036)		(0.010)	(0.046)		(0.018)	(0.039)
Share white and Asian	0.178		0.761	-0.844		0.350	-0.161
	(0.127)		(0.025)	(0.233)		(0.043)	(0.123)
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 1. Standard errors are reported in parentheses.

Table A5. Projections of School Quality and School Quality on Share Disadvantaged

		Test score levels			Test score progress		
	VA projection (derived)	VA projection (derived)	Rating projection (OLS)	VA projection (derived)	VA projection (derived)	Rating projection (OLS)	VA 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. NYC				
Predictors							
Test score levels		0.219		0.480			
		(0.057)		(0.069)			
Test score progress					0.779		0.797
					(0.041)		(0.041)
Screened school dummy	-0.049		0.059	-0.077		-0.041	-0.016
	(0.037)		(0.012)	(0.037)		(0.016)	(0.034)
Share non-FRPL	-0.022		0.636	-0.327		0.143	-0.136
	(0.054)		(0.017)	(0.063)		(0.024)	(0.050)
N (school-year)				1501			
			Panel B. Denver				
Predictors							
Test score levels		0.521		1.23			
		(0.136)		(0.255)			
Test score progress					0.915		0.957
					(0.095)		(0.114)
Charter school dummy	0.118		0.053	0.052		0.107	0.016
	(0.036)		(0.010)	(0.041)		(0.018)	(0.039)
Share non-FRPL	0.165		0.660	-0.649		0.288	-0.111
	(0.113)		(0.026)	(0.203)		(0.038)	(0.111)
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 eligible for a free or reduced-price lunch. The models and derivation procedure used to compute these estimates are as the estimates in Table 1. Standard errors are reported in parentheses.

Table A6. IV VAM Long Regressions

	Over-identified (school assignment instruments)		Just-identified (offered mediator instruments)	
	NYC	Denver	NYC	Denver
	(1)	(2)	(3)	(4)
Mediators				
Test score levels	-0.211	0.175	-0.217	-0.038
	(0.064)	(0.289)	(0.077)	(0.466)
Test score progress	-0.015	0.382	-0.042	0.313
	(0.101)	(0.198)	(0.116)	(0.232)
RC VAM	1.07	0.612	1.13	0.680
	(0.111)	(0.176)	(0.134)	(0.204)
Screened school dummy	0.002		0.014	
	(0.033)		(0.036)	
Charter school dummy		-0.027		0.011
		(0.043)		(0.062)
Share white	0.005	-0.288	0.030	-0.061
	(0.064)	(0.243)	(0.079)	(0.375)
First-stage F	11.8	11.7	294	22.1
N	46,099	8,100	46,099	8,100

Notes: This table reports IV VAM parameter estimates. These estimates are used to obtain the estimates reported in Table 1. The set of listed mediators is treated as endogenous. Columns 1 and 2 use individual school assignment offer dummies as instruments. Columns 3 and 4 use values of the mediator at the offered school as instruments. Models are estimated using 2SLS. RC VAM estimates come from individual student test scores on school enrollment dummies, baseline demographic and lagged score controls, and assignment risk. Ratings are demeaned and scaled to have variance matching that of β_j across schools in the district. Standard errors are reported in parentheses.