Blueprint Labs

Discussion Paper #2025.02

Putting School Surveys to the Test

Joshua Angrist
Peter Hull
Russell Legate-Yang
Parag A. Pathak
Christopher R. Walters

April 2025

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

Putting School Surveys to the Test*

Joshua Angrist† Peter Hull‡ Russell Legate-Yang§
Parag A. Pathak¶ Christopher R. Walters‖

April 4, 2025

Abstract

School districts increasingly gauge school quality with surveys that ask about school climate and student engagement. We use data from New York City's middle and high schools to compare the long-run predictive validity of surveys with that of conventional test score value-added models (VAMs). Our analysis leverages the New York school match, which includes an element of random assignment, to validate a wide range of school quality estimates. We contrast the predictiveness of survey- and test-based measures for school effects on consequential outcomes related to high school graduation and college enrollment. Survey data generate better predictions of school impacts on high school graduation than test scores. But school effects on advanced high school diplomas and college attainment are better predicted by test score VAMs than surveys. We quantify the practical value of test-based and survey-based school quality measures by simulating the effects of access to one or both types of information for parents. Parents interested in boosting their children's college attainment benefit more from test score value-added than from survey data.

*Financial support from the Gates Foundation (grant INV-043521) is gratefully acknowledged. Views expressed here are those of the authors and do not necessarily reflect positions or policies of the foundation. Legate-Yang also gratefully acknowledges financial support from the Paul & Daisy Soros Fellowship. We thank Eryn Heying, Talia Gerstle, and Niamh McLoughlin for invaluable administrative support. We also thank Eric Ashton and the staff of the New York City Public Schools (NYCPS) as well as Peter Blair and Caroline Hoxby and participants in the Fall 2024 NBER Education meeting for insightful comments. Fatima Djalalova provided expert research assistance. The research described here was carried out under data-use agreements between MIT and the New York City school district. The IRB of record is MIT COUHES, under protocol 1904808765.

†MIT and NBER. Email: angrist@mit.edu
‡Brown University and NBER. Email: peter_hull@brown.edu
§MIT. Email: rcly@mit.edu
¶MIT and NBER. Email: ppathak@mit.edu
‖University of California, Berkeley and NBER. Email: crwalters@econ.berkeley.edu

1 Introduction

For decades, school districts and states have relied on standardized tests to measure school performance—but this has started to change. The 2015 *Every Student Succeeds Act* requires states to include non-academic factors in federally-mandated school accountability reports (Jordan and Hamilton, 2020). At the same time, social scientists and educators have argued for the importance of non-cognitive development for students' later life outcomes and enlisted school surveys to implement these ideas (see, e.g., Dweck, 2006; Duckworth et al., 2007; Cunha and Heckman, 2007; Jackson et al., 2020; Hart et al., 2020; Schneider et al., 2021; Deming, 2023). Responding to such currents, school districts increasingly disseminate and highlight survey-based measures of school climate and student engagement. Fourteen states include or plan to include data from climate surveys in published accountability statistics (Learning Policy Institute, 2024). Demand for school surveys is also evident in the growth of firms like Panorama Education, a survey contractor that serves 2,000 districts including 50 of the 100 largest, covering 25% of U.S. K-12 students (Panorama Education, 2025).

The emerging emphasis on survey-based measures of school performance has often (and sometimes intentionally) come at the expense of test-based accountability schemes. New York City is a leading example of this trend: in 2013, it abandoned a largely test-based accountability regime established under outgoing Mayor Michael Bloomberg. In the Bloomberg era, many schools—especially those serving low income students—struggled to earn high grades on the city's "school report cards."[1] In 2014, the mayoral administration of Bill de Blasio introduced an accountability scheme that prioritizes survey data over standardized tests (Taylor, 2014). Based in part on a redesigned school climate survey, the new measures are meant to provide a more complete picture of school performance (Merrill et al., 2018). In practice, the new ratings feature variables such as the proportion of students indicating that they "learn a lot from feedback on their work." Critics of the new approach argue that such reports are inherently vague and likely to be inconsistent (Wall, 2014). Similar debates on the merits of self-assessed learning measures feature in analyses of college students' course evaluations (Hoffmann and Oreopoulos, 2009; Weinberg et al., 2010; Carrell and West, 2010; Braga et al., 2014; Stark and Freishtatd, 2014).

The New York City public school survey is one of the most ambitious efforts of its kind. This survey design draws on research identifying aspects of the school environment that are widely seen as related to school improvement (Merrill et al., 2018). The New York survey has intellectual roots in a similar effort based in the Chicago public schools (Bryk et al.,

[1]Bloomberg-era report cards were 85 percent test-score based (Hernández, 2013). Corradini (2024) examines the impact of school report cards and other performance measures on student application patterns.

2010). Much of the information on the New York City Public Schools school ratings website, presented to families in the form of "school quality snapshots," originates in these surveys. School quality snapshots present test-based measures as well, though the details are found on a separate "school performance dashboard" that parents are less likely to encounter.[2]

This paper evaluates and contrasts the predictive validity of survey-based and test-based measures of school quality in New York. Building on Angrist et al. (2024a), our analysis leverages the random assignment embedded in New York's middle and high school matches to validate causal school effects in a value-added model (VAM) framework. Specifically, we use the New York match to estimate and validate effects of individual school attendance on achievement as measured by test scores, a summary of survey-based performance indicators, and consequential longer-term outcomes related to high school graduation and college attendance. The internal validity of test- and survey-based measures is evaluated by asking whether students randomly assigned to schools with better performance measures see higher achievement and report higher school performance as a result.

After validating measures of school quality based on tests, surveys, and longer-term outcomes, we study relationships between these. Our analysis is in part motivated by Bryk et al. (2010), which makes a case for school surveys based on correlations between surveys and VAM-like measures. New York's school survey was inspired by these results.[3] Parents and students may value school school climate and a high level of student engagement regardless of achievement value-added. At the same time, positive survey reviews need not generate higher human capital or increase post-secondary attainment. Our analysis characterizes links between test-VAMs, survey assessments, and longer-term outcomes by regressing high school graduation and college enrollment value-added on survey and test score measures of school quality. We also summarize the predictive value of alternative measures of school performance in a decision-theoretic framework, in which applicants pick schools so as to maximize expected college attendance based on different portfolios of school quality information.

Survey assessments are found to generate better predictions of school effects on high school graduation rates than test score VAM, especially for high schools. A high school rated one standard deviation higher on surveys is estimated to increase on-time high school diploma receipt by 11 percentage points, on average. However, test score VAM tops survey information as a guide to longer-term school quality. In particular, survey reviews of both middle and high schools predict advanced diploma and college attainment effects relatively weakly. Test score value-added, by contrast, strongly predicts school impacts on advanced

[2]See, for example, the quality snapshot for Beacon HS and the corresponding performance dashboard.

[3]A report by the research team that helped design the NYC survey writes: "Bryk et al. (2010) showed that [the 'Five Essentials' survey measures] predicted improvements in students' value-added scores. Therefore, we chose to include all of these in the new NYC Survey" (Merrill et al., 2018, p. 5).

diplomas and college attainment at both the middle and high school levels. For instance, high schools boosting test scores by one standard deviation are estimated to increase on-time college enrollment by 11 percentage points. Simulations of school choice under uncertainty show that test score information achieves around 30% of the theoretical payoff to perfect quality information. Survey information does little to improve on this.

The remainder of the paper is structured as follows. The next section describes the New York City schools setting as well as the survey and test score data underlying our analysis. Section 3 details the econometric framework used to estimate and validate VAMs for school effects on test scores, surveys, and longer-run outcomes related to graduation and college enrollment. This section also explains how we assess the predictive power of survey and test measures for school effects on longer-run outcomes. Our analysis is predicated on the notion that parents and educators aim to use published school quality information to gauge impacts on important longer-run outcomes. Section 5 formalizes this view by simulating a school choice problem in which parents use test and survey information to maximize expected college attendance for their children. Section 6 concludes.

2 Setting and Data

2.1 Data

Our analysis uses administrative data on public school students provided by the New York City Public Schools (NYCPS). The core sample consists of fifth-grade applicants to New York City middle schools for the 2015-2016 school year and eighth-grade applicants to New York City high schools for the 2012-2013 through 2016-2017 school years. The data include each applicant's rankings over schools, priorities at each school program, and offers made by the centralized assignment system. We also have data on students' demographics, school enrollment, and achievement outcomes. Achievement outcomes consist of New York State standardized test outcomes in grade 6 and scores from the first Regents math exam taken in high school. The latter exam is the Algebra I Regents for most high schoolers, though for some their first high school Regents math test is Geometry. Test scores are standardized to be mean zero and standard deviation one among all test takers for a given subject in each year. School survey information comes from New York's school climate surveys, from grade 9 responses in high school and grade 6 responses in middle school.

We link estimates of school quality based on tests and surveys to longer-run high school graduation and college attendance outcomes. Most students graduating from NYC high schools earn a Regents diploma, but a large minority graduate with the more demanding advanced Regents diploma. The basic Regents diploma requires students to earn minimum scores on five Regents exams along with a core set of course credits, while an advanced diploma requires passing scores on nine tests and additional course credits. The stakes for advanced diploma receipt include qualification for direct admission to many of New York State's SUNY campuses (Zimmerman, 2024). Information on college attendance comes from a match of New York City high school graduates to the National Student Clearinghouse (NSC) data. Our analysis focuses on two college outcomes: on-time enrollment, defined as an indicator for enrollment in any college (two- or four-year) by the fall of a student's expected graduation year based on her middle or high school application cohort; and 2-year persistence, an indicator which equals one for students who enroll on-time and maintain enrollment for five consecutive semesters or graduate within that period. See the Data Appendix for details on data sources and outcome construction.

2.2 School Assignment in New York City

New York uses the student-proposing deferred acceptance (DA) algorithm with a single lottery tie-breaker to assign students to schools in both middle and high school. Entry-grade applicants in this system submit rank-ordered lists of up to 12 academic programs. A program defines a course of study at a particular school (schools may have more than one). Students are assigned priorities at each ranked school based on criteria like sibling status and neighborhood of residence. Ties among equal-priority applicants are broken with a mix of lottery and non-lottery tie-breakers. At the city's "unscreened" schools, the tie-breaker is a random lottery number. At "screened" schools, the non-lottery tie-breakers can include past test scores, interview rankings, attendance, or grades. The centralized match collects all applicants' rank-ordered preference lists, priorities, and tie-breakers, and executes the DA algorithm to generate a single school assignment for each student.

Our analysis leverages information from the centralized match to construct and validate measures of school effectiveness. We use methods detailed in Abdulkadiroğlu et al. (2017, 2022) to compute each applicant's *DA propensity score* at each school, defined as the probability (or *risk*) that the student is assigned to the school over repeated runs of the assignment algorithm. Our analysis aggregates program-level assignment risk for multiple programs in a school to create school-level propensity scores. Assignment risk equals zero for a student who has no chance to be seated at a school (e.g. because she did not rank it), equals one for a student who is guaranteed a seat (e.g. because she ranks the school first and it has fewer applicants than seats), and is strictly between zero and one for applicants at over-subscribed schools whose assignments are partly determined by tie-breaking. As in Abdulkadiroğlu et al.

(2022), we incorporate non-lottery tie-breaking by focusing on students with tie-breakers in a small bandwidth of screened school admission cutoffs, treating assignment as a local random lottery within the bandwidth. The upshot is that conditioning on the DA propensity score yields a stratified randomized trial, with strata defined by score values.[4]

2.3 New York School Surveys

The NYC School Survey has been fielded annually since 2006 to students in grades 6-12, their families, and school teachers and staff. Students in our sample period complete paper or digital surveys during a dedicated class period scheduled by their school between February and April. The survey captures student and family identifiers as well as anonymized information from educators. Our data covers survey responses since 2014, when the survey was redesigned following the election of a new mayor. The 2014 redesign was inspired by surveys developed by the University of Chicago Consortium of School Research (Bryk et al., 2010; Merrill et al., 2018). Since 2014, student survey response rates have ranged from around 76% in high school to 91% in middle school. Because parent response rates are much lower (around 35% for HS and 57% for MS), we focus on student survey data. Our analysis ignores educator surveys since these likely reflect factors—such as seniority—governing teacher school assignment as much as school quality.[5]

We construct a survey index for middle school and high school students by averaging standardized item responses in 6th grade for middle school and mostly 9th grade for high school.[6] Survey item responses are captured on a categorical 1-4 scale that measures agreement with the stated question. We follow New York's public survey reports and recode items so that higher responses are more favorable. Survey timing corresponds to typical testing times for our test score outcomes. We drop students who answer no questions. As with test scores, we standardize the survey index to have mean zero and standard deviation one across NYC students each year. See the Data Appendix for more details.

[4] Appendix Tables A1.A, A1.B, and A2 find little evidence of statistical imbalance or differential attrition in the natural experiments generated by centralized assignment for middle and high school match applicants.

[5] Appendix Section B examines the predictive validity of independent experts hired by the district to rate schools. This analysis suggests expert reviews respond more to student family background and ability than to schools' causal effects on long-term outcomes.

[6] Jackson (2018) uses principal component analysis to construct a behavior index from a combination of grades, on-time grade progression, absences, and suspensions (see Heckman et al. (2006); Jackson et al. (2020) for related approaches). Table A7 shows that a coding scheme based on the first principal component of survey responses generates results similar to those reported in the text.

2.4 Descriptive Statistics and Assessment Correlations

Table 1 describes the sample of students and schools. The sample with risk, for which statistics are reported in columns 2 and 4 of the table, includes students for whom school assignment has some randomness. This sample is used to validate school quality measures. The high school risk sample includes students at all high schools while the middle school risk sample covers 584 out of 616 schools. Most schools with risk are unscreened, i.e., they use lottery tie-breaking. Proportionally fewer middle schools are screened than high schools.

About three-quarters of New York's high school students are Black or Hispanic, while nearly 80% qualify for a subsidized school meal—a traditional indicator of poverty. The demographic makeup of students, the distribution of baseline scores, and the enrollment patterns of students with risk are demographically similar to the full sample. Middle school students are similar to high school students in most respects but are somewhat less likely to be Hispanic or Black.

Figure 1 motivates our empirical analysis by plotting school-level relationships between college enrollment rates and either test score or survey averages. Across both high schools and middle schools, a one-decile increase in test scores is associated with around a 4.5 percentage point increase in college enrollment. The corresponding increase for survey averages is smaller: 2.1 percentage points for high schools and 1.0 percentage points for middle schools.

Test- and survey-based school quality measures are correlated but distinct. This is documented in Figure 2 in the form of heatmaps describing correlations between a broad set of achievement and survey measures. Survey item responses are organized into categories according to New York's "Framework for Great Schools" school performance scheme (see the Data Appendix for details). As can be seen in the bottom right block of the figure, survey assessments for different categories are highly correlated. Test, diploma, and college attainment levels, marked in the upper left quadrant, are also highly correlated. The two-block correlation structure apparent in the figure suggests that surveys capture aspects of school quality distinct from other measures. This correlation, however, may reflect patterns of selection bias as well as causal effects. We next develop an econometric framework that aims to isolate relationships between school causal effects, however measured.

3 Econometric Framework

3.1 Multivariate School Value-Added

A multivariate value-added framework allows for school effects on multiple short- and long-term student outcomes.[7] In this framework, a collection of outcomes Y_i^k, for $k \in \{1, \ldots, K\}$, characterizes student i's educational trajectory. Longer-run outcomes like high school graduation and college enrollment are important education policy targets. Short-run outcomes like standardized test scores and survey responses are of interest in part because they predict later, more consequential outcomes. Achievement as gauged through test scores might also be indicative of valuable skills like critical thinking. Although we don't usually think of schools as aiming for good reviews, survey reports of positive in-school experiences might signal increased learning or improved hard-to-measure skills like grit.

A constant-effects causal model describes student i's potential values for each outcome k if he or she were to attend school $j \in \{1, \ldots, J\}$:

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

where β_j^k is school j's value-added for outcome k. The vector of value-added effects $\beta_j = (\beta_j^1, ..., \beta_j^K)'$ characterizes school j's contributions to student development. Error term ε_i^k captures non-school factors that influence student i's outcome on dimension k, such as family background, motivation, and ability. Because $\beta_j^k = E[Y_{ij}^k]$ is defined as the population mean of potential outcome k at school j, the error term ε_i^k has population mean zero. The constant effects assumption, reflected in the additive structure of equation (1), abstracts from heterogeneity in school effects across students. But the framework outlined here can be readily extended to accommodate idiosyncratic match quality unrelated to school enrollments. We explore effect heterogeneity by observable student characteristics below, finding minimal impacts on our empirical analysis.

The observed outcome for student i is the potential outcome corresponding to her enrolled school. Letting D_{ij} denote an indicator equal to one if i enrolls at school j, observed outcomes are linked with potentials by:

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

School enrollment is not randomly assigned, and students choosing different schools are likely

[7] See Jackson (2018), Abdulkadiroğlu et al. (2020), Beuermann et al. (2022), and Rose et al. (2022) for recent related analyses of multi-dimensional teacher and school effects.

to differ in many ways. This can lead to selection bias in uncontrolled comparisons of mean outcomes across schools. Selection bias in a comparison of average outcomes at schools j and j' is characterized by writing:

$$
\begin{aligned}
E[Y_i^k|D_{ij}=1] - E[Y_i^k|D_{ij'}=1] &= \beta_j^k - \beta_{j'}^k + \{E[\varepsilon_i^k|D_{ij}=1] - E[\varepsilon_i^k|D_{ij'}=1]\} \\
&\neq \beta_j^k - \beta_{j'}^k. \qquad (3)
\end{aligned}
$$

We mitigate selection bias by controlling for variables that generate correlation between school attendance and potential outcomes. Let X_i denote a vector of control variables (demeaned so that $E[X_i] = 0$), and let γ^k denote the coefficient from a projection of ε_i^k on X_i. Substituting this projection into equation (2) yields

$$
Y_i^k = \sum_j D_{ij}\beta_j^k + X_i'\gamma^k + e_i^k, \qquad (4)
$$

where $E[e_i^k] = E[e_i^k X_i] = 0$ by definition of γ^k. Our value-added modeling approach is predicated on the assumption that $E[D_{ij}e_i^k] = 0$ for all j and k. This selection-on-observables restriction implies that equation (4) coincides with an ordinary least squares (OLS) regression of Y_i^k on school indicators with controls for X_i. Importantly, given conditional random assignment of some students to schools in New York, this assumption is testable.

Control vector X_i includes student demographic characteristics, lagged test scores, and assignment risk controls (also called propensity scores) giving the probability student i is seated at school s for each student and school. The inclusion of assignment risk controls makes equation (4) a *risk-controlled value-added model* (RC VAM). In a study of value-added in New York middle and high schools and Denver middle schools, Angrist et al. (2024a) validates RC VAM estimates by showing these estimates predict the impact of random school assignment on test scores. This suggests RC VAM estimates offer an approximately unbiased gauge of school effects. Section 4.1 reports results of similar tests for the NYC samples and outcomes used here, likewise validating RC VAM estimates as a measure of school causal effects.

3.2 Relationships Between School Effects

Our analysis focuses on relationships between school effects across multiple student outcomes. These relationships are summarized by a list of covariance parameters:

$$
\sigma_{km} = \frac{1}{J-1}\sum_{j=1}^{J}(\beta_j^k - \mu^k)(\beta_j^m - \mu^m), \qquad (5)
$$

where $\mu^k = (1/J)\sum_j \beta_j^k$ is the average value-added for outcome k. For a single outcome (setting $m = k$), this expression gives the variance of value-added in the finite population of schools in New York City—a measure of the variability of school quality on dimension k. For different outcomes ($m \neq k$), σ_{km} is the covariance between value-added parameters across equations, revealing whether schools that boost outcome k also tend to boost outcome m.

VAM variance and covariance parameters can be expressed compactly as quadratic forms in a vector of unknown value-added parameters. Let $\beta = (\beta_1^1, \beta_2^1, ..., \beta_J^1, \beta_1^2, ..., \beta_J^K)'$ denote the $JK \times 1$ vector collecting value-added coefficients across all schools and outcomes. Then

$$\sigma_{km} = \beta' A_{km} \beta, \tag{6}$$

where A_{km} is a $JK \times JK$ matrix composed of $J \times J$ blocks of zeros at all positions except its (k, m)th block, which equals $(J-1)^{-1}(I_J - J^{-1}\iota_J\iota_J')$, with I_J the $J \times J$ identity matrix and ι_J a $J \times 1$ vector of 1's.

We construct estimates of these quadratic forms using VAM estimates and their sampling variances and covariances.[8] Specifically, let $\hat{\beta}$ denote the $JK \times 1$ vector of value-added estimates generated from joint estimation of system (4) for all K outcomes. These estimates are assumed to satisfy $E[\hat{\beta}] = \beta$ and have sampling variance described by a $JK \times JK$ matrix $V = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$, which can be well-approximated based on standard asymptotic theory.

A plug-in estimator of σ_{km} substitutes an OLS estimate $\hat{\beta}$ for β in equation (6). Because the quadratic form is a nonlinear function of β, the resulting variance estimate is biased. Intuitively, the variance of value-added coefficients for each outcome is inflated due to sampling error in estimated VAM, while the covariance between value-added estimates for different outcomes is biased due to correlation in the underlying student-level outcomes. These biases are characterized by

$$E[\hat{\beta}' A_{km} \hat{\beta}] = \beta' A_{km} \beta + tr\left(A_{km} V\right). \tag{7}$$

This formula suggests bias-corrected estimators for VAM covariance components of the form,

$$\hat{\sigma}_{km} = \hat{\beta}' A_{km} \hat{\beta} - tr\left(A_{km} \hat{V}\right), \tag{8}$$

where $\hat{V}$ is a White (1980) heteroskedasticity-robust estimate of V.

Bias-corrected covariance estimates are used to construct coefficients from bivariate and multivariate regressions relating VAM parameters for various outcomes. For example, the slope coefficient from a bivariate regression of β_j^k on β_j^m is estimated as $\hat{\sigma}_{km}/\hat{\sigma}_{mm}$. Likewise, estimates of multivariate regression coefficients are constructed using the set of $\hat{\sigma}_{km}$'s and the

[8]See Walters (2024) for a detailed discussion of this bias-corrected variance estimation approach.

relevant regression anatomy formulas. These coefficients allow us to compare the predictive value of test score value-added and surveys for school effects on longer-run outcomes like college attendance, either one-at-a-time or jointly. Standard errors for the estimates of interest are computed using the delta method.[9]

4 Results

4.1 Validating VAMs

We assess VAM validity using lottery-based bias tests of the sort developed in Angrist et al. (2016) and Angrist et al. (2017). These tests are based on estimates of the equation

$$Y_i^k = \alpha^k + \varphi^k \hat{\beta}^k_{d(i)} + p_i'\delta^k + \eta_i^k, \tag{9}$$

where $\hat{\beta}^k_{d(i)} = \sum_j D_{ij}\hat{\beta}^k_j$ is an estimate of value-added on outcome k for student i's enrolled school $d(i)$ and $p_i = (p_{i1}, ..., p_{iL})'$ is a vector of assignment propensity scores for a set of L schools. This set is limited to oversubscribed schools, defined as those with non-degenerate assignment risk (i.e. $p_{i\ell}$ strictly greater than zero and less than one) for some students i. The validation sample includes students with non-degenerate risk for at least one school.

Equation (9) is estimated by instrumenting enrolled value-added, $\hat{\beta}_{d(i)}$, with a set of L offer indicators,[10] $Z_{i\ell}$, for oversubscribed schools using two-stage least squares (2SLS). Offer dummies are randomly assigned conditional on the risk controls p_i. When VAM is unbiased, random offers that shift students across schools should cause outcome Y_i^k to increase one-for-one with estimated enrolled value-added $\hat{\beta}^k_{d(i)}$, implying the 2SLS *forecast coefficient* φ^k should equal one. The deviation of the 2SLS estimate from unity therefore provides a sense of average VAM bias, or "forecast bias." With multiple oversubscribed schools ($L > 1$) φ is overidentified, yielding a 2SLS overidentification test that checks whether the forecast coefficients generated by each offer instrument are equal. An omnibus test for VAM bias checks all implications of a valid VAM by combining the forecast and overidentification tests into a test with L degrees of freedom (Angrist et al., 2016, 2017).

Table 2 reports validity test results for unadjusted levels and for VAMs adjusting for

[9]The delta method sampling variance of $\hat{\sigma}_{km}$ is also a quadratic form in $\hat{\beta}$. This is bias-corrected using $\widehat{Var}(\hat{\sigma}_{km}) = \hat{\beta}'\tilde{A}_{km}\hat{\beta} - tr(\tilde{A}_{km}\hat{V})$, where $\tilde{A}_{km} = (A_{km} + A'_{km})\hat{V}(A_{km} + A'_{km})$. When school parameters β_j are viewed as random draws from a school quality distribution rather than as fixed parameters, addition of $(\hat{\sigma}^2_{km} + \hat{\sigma}_{kk}\hat{\sigma}_{mm})/(J-1)$ to the formula for the estimated variance of $\hat{\sigma}_{km}$ captures the resulting extra uncertainty. In practice, this random-effects-style correction matters little.

[10]In practice, offers are classified into twenty bins defined by ventiles of the distribution of estimated VAM.

observed characteristics. Estimates for high schools appear in Panel A and estimates for middle schools appear in Panel B. VAM validity tests for survey responses, reported in columns 3 and 4, show little bias in unadjusted survey levels. 2SLS forecast coefficient estimates for survey levels are close to one, while the omnibus test fails to reject the null hypothesis of unbiasedness with no controls. This suggests that simple school survey averages provide unbiased estimates of the causal effects of schools on students' survey responses.

Although survey levels appear to be unbiased, validity test results for other outcomes reveal substantial selection bias in naive comparisons. Forecast coefficients for these measures range from 0.13 to 0.51, and the associated omnibus test statistics show decisive rejections. In contrast, estimates in columns 2, 6, and 8 of Table 2 suggest that RC VAM estimates align with school causal effects for most of the outcomes studied here. 2SLS forecast coefficient estimates are statistically indistinguishable from one for test and college outcomes in both middle and high school and for the graduation outcomes in middle school, while the corresponding omnibus tests fail to reject the null of RC VAM validity for all but the high school test outcome. The sole exception to this pattern is the test for basic Regents Diploma effects in high school, where the forecast coefficient estimate of $\hat{\varphi} = 0.80$ is significantly different from one and the omnibus test indicates a decisive rejection. Value-added estimates for this outcome should therefore be interpreted more cautiously than results for test scores, surveys, and college attendance.[11]

Why are raw survey responses a good guide to causal survey effects whereas uncontrolled comparisons of other outcomes are misleading? Appendix Table A4 investigates this by summarizing relationships between test and survey outcomes, lagged values of test and survey variables, and school enrollment. Baseline (middle school) test scores are strong predictors of high school test scores but don't predict high school survey responses. High school survey responses are less correlated with baseline survey responses, suggesting less scope for selection bias in these. Baseline test scores are also highly variable across high schools, while baseline survey responses are more balanced: adjusted r-squareds from regressions of baseline tests and surveys on high school indicators equal 0.26 and 0.03, respectively. These relationships explain why survey measures are relatively insensitive to baseline test and survey controls.[12] Evidently, there is little sorting of students to schools on the basis of potential survey responses, while sorting on potential achievement is substantial. In this context, it's also worth noting that while interested parents might find their way to test VAM details on

[11]VAM bias tests for other outcomes appear in Appendix Tables A3.A and A3.B.

[12]Recall from the omitted variable bias formula that the bias from omitted controls depends on the long-regression control coefficient multiplied by the coefficient from a regression of the omitted variable on the included regressor. Since baseline test scores do not predict survey outcomes and baseline survey responses are balanced across schools, estimates of school effects on survey outcomes are insensitive to these controls.

the NYCPS's school performance dashboard, the NYCPS reports only unadjusted average survey responses. These factors motivate our juxtaposition of unadjusted survey levels with RC VAM estimates for other outcomes.

Appendix Table A5 shows bias-corrected estimates of the standard deviation of value-added based on the VAM considered in each column, $\sqrt{\hat{\sigma}_{kk}}$. Consistent with the view that RC VAM eliminates the substantial selection bias in achievement levels, RC VAM generates lower estimates for the standard deviations of school quality than do the levels measures (i.e. some of the dispersion in school average outcomes is due to selection bias rather than school quality). Nonetheless, RC VAM estimates still point to substantial variation in school quality across all outcomes for both middle and high schools. Across high schools, for example, a one-standard deviation increase in school quality boosts test scores by 0.23 standard deviations of the student-level distribution, survey responses by 0.29 student-level standard deviations, and on-time enrollment at any college by 7 percentage points.

4.2 Contrasting Survey and Test VAM: Regression Prediction

Survey measures predict high school impacts on timely Regents diploma receipt more strongly than does test score value-added. This can be seen in Table 3, which uses bias-corrected variance and covariance estimates ($\hat{\sigma}_{km}$) to summarize relationships between test value-added, survey averages, and longer-run outcome value-added. To facilitate comparisons, these bias-corrected regression estimates were constructed using test score and survey measures standardized to have variance one across students. Regression models can therefore be interpreted as the gain in, e.g., college value-added in percentage points yielded by a school that boosts test value-added or survey means by one standard deviation in the student-level distribution of raw tests or surveys.

Among NYC high schools, a one standard deviation increase in average survey responses is associated with an 11 percentage point increase in school effects on Regents diploma receipt. Test VAM yields about 9 points on the same scale. These estimates appear in the first two columns of Panel A in Table 3. In a multivariate model that includes both test and survey predictors, estimates of which are reported in column 3, the coefficient on test VAM is statistically indistinguishable from zero. This finding should be qualified somewhat by the fact that RC VAM for high school effects on timely Regents diploma receipt fails a validity test (see Table 2). It's noteworthy, however, that the survey advantage in this context is also apparent for middle schools, though the magnitude of the survey coefficient is diminished relative to that for high schools. Moreover, middle school test-score VAM has no predictive power for middle school effects on Regents diploma receipt.

Looking at more demanding high school graduation and college outcomes, test score value-added beats survey levels as a guide to school quality. Specifically, columns 4-6 of Panel A in Table 3 show that surveys are unrelated to value-added for advanced diploma receipt, while a one standard deviation improvement in test value-added predicts a roughly 11 percentage point advanced diploma gain with or without control for surveys. Estimates of middle school effects on advanced diplomas, reported in Panel B, also favor test VAM over surveys, though magnitudes are again diminished relative to those for high schools.

Regression predictions for college outcomes likewise favor test VAM over surveys. Table 4, formatted like Table 3, reports test and survey effects on college value-added. The first column of Panel A shows a precisely estimated 11 percentage point gain in college enrollment for high schools. The corresponding survey coefficient, reported in column 2 of the table, is only around 6 percentage points. A multivariate model including both test value-added and survey measures also generates a strong test effect with an even smaller survey coefficient. This suggests that the modest predictive value seen for surveys in the bivariate model reflects correlation between surveys and test VAM.

Results for high school effects on college persistence, reported in columns 4-6 of Panel A, are much like those for college enrollment in that they clearly favor test value-added over surveys. Finally, estimates of middle school effects on college enrollment, reported in Panel B of Table 4, generate small survey coefficients that are not significantly different from zero. At the same time, effects of middle school test value-added on college enrollment value-added hover around 17 percentage points.

4.3 Robustness: Outcome and VAM Alternatives

The predictive power of test VAM for demanding graduation and college attainment outcomes is even more pronounced for VAM constructed from SAT math and ELA scores rather than Regents Algebra 1 and Geometry scores. Estimates using SAT VAM, given in Appendix Table A6, show test effects of around 17 percentage points for advanced diplomas and college enrollment and around 15 percentage points for college persistence. The corresponding effects for survey levels, measured with grade 11 responses (the typical SAT taking year), are small and insignificant or nearly so. It is also worth noting that the gain in predictive power from Regents math to SAT VAM is largest for college persistence, an outcome less mechanically linked to SAT scores than college enrollment.[13]

Our survey measure averages item responses to all survey questions. Appendix Table A7 shows that replacing this average with a first principal component for survey responses or

[13]The SAT taking rate, sometimes thought to indicate college aspirations, is 68% in the samples here. Three of the five high school cohorts were offered free in-school SAT testing (Veiga, 2017).

survey category averages (similar to those reported in New York's accountability framework) leaves results unchanged.

Academic research on college attainment often focuses on 4-year college enrollment as a marker of success. The estimates in Appendix Table A8, computed for 4-year college enrollment in a setup similar to that used to construct the estimates in Table 4, likewise favor test value-added over surveys. Finally, Appendix Table A9 reports estimates for high schools replacing on-time high school graduation and college enrollment outcomes with comparable variables measured six years after high school entry (data constraints preclude a similar analysis for middle schools). Again, these estimates favor surveys for Regents diplomas and test value-added for college.

As noted in Section 4.1, validity tests reject the null hypothesis of unbiasedness for high school Regents diploma VAMs. Appendix Tables A10 and A11 therefore report results constructed using an alternative instrumental variables value-added model (IV VAM) that relies only on conditional random assignment of students to schools to identify causal school effects. The IV VAM model (detailed in Angrist et al., 2024a) uses conditionally random school offers to instrument for a set of mediators—in this case, test and survey measures—in an equation for student-level longer-term outcomes. Replacing RC VAM estimates with IV VAM estimates in the procedure used to construct Tables 3 and 4 generates a similar constellation of findings favoring surveys for school effects on Regents diploma receipt and test value-added otherwise.

We also recognize that districts without centralized assignment or decentralized controls cannot construct the risk controls underpinning RC VAM or the instrumental variables underpinning IV VAM. Appendix Table A12 therefore looks at the predictive value of conventional VAM estimates that rely solely on widely-available demographic and lagged test score controls as well as the simplest VAM estimator, often referred to by districts as "progress," equal to school-average achievement test growth. Patterns here are again similar.

5 VAM Without Regrets

The previous section contrasts the predictive value of test VAM and survey levels for longer-term value-added, measured in standardized units. From the perspective of a household aiming to promote longer-term student success, the key question is whether using these measures as a guide to school enrollment leads to meaningful improvements in high school graduation and college attendance. A decision-theoretic framework quantifies the value of different types of school information for this purpose.

Suppose a parent seeks to promote college attendance (outcome K) for his or her child, and has access to a list of estimates that may include survey results, test score value-added estimates, or both. Let $\mathcal{P}$ denote a parent's information portfolio, and let $\mathcal{C} \subseteq \{1, ..., J\}$ denote the corresponding school choice set. Parents aim to choose school j from $\mathcal{C}$ with the highest college attendance potential outcome, Y_{ij}^K. Model (1) implies that the highest potential outcome corresponds to the school with the highest college value-added β_j^K. Value-added parameters are unknown, so parents choose schools to maximize expected college value-added given information $\mathcal{P}$. This yields expected outcome

$$U(\mathcal{P}; \mathcal{C}) = E\left[\max_{j \in \mathcal{C}} E[\beta_j^K | \mathcal{P}]\right]. \tag{10}$$

The outer expectation in equation (10) averages over $\mathcal{P}$ as well as the value-added vectors $\beta_j = (\beta_j^1, ..., \beta_j^K)'$, treating the latter as random effects drawn from a multivariate distribution defined in the population of schools.

Regret contrasts the expected outcome achieved by a parent equipped with estimated VAM or survey information with the value achieved by an *oracle* that knows the vector of value-added parameters for all schools. The oracle's outcome with choice set $\mathcal{C}$ is given by $U(\beta; \mathcal{C})$. The *regret* of a parent with information portfolio $\mathcal{P}$ is defined as the difference in payoffs relative to this oracle benchmark, given by

$$\mathcal{R}(\mathcal{P}; \mathcal{C}) = U(\beta; \mathcal{C}) - U(\mathcal{P}; \mathcal{C}). \tag{11}$$

The oracle's expected outcome provides an upper bound on the gains afforded by school quality information, measured in units of expected college attendance. For each information portfolio $\mathcal{P}$, we compute regret $\mathcal{R}(\mathcal{P}; \mathcal{C})$ as well as the proportional reduction in regret relative to the outcome for an uninformed parent, given by $1 - [\mathcal{R}(\mathcal{P}; \mathcal{C}) / \mathcal{R}(\emptyset; \mathcal{C})]$.

Regret parameters are calculated via a school choice simulation with value-added distributions calibrated to match our empirical estimates. The simulation draws true value-added vectors β_j for each school from a joint normal distribution with covariance matrix populated by bias-corrected estimates of the covariance parameters σ_{km}. We then draw VAM estimates $\hat{\beta}_j = \beta_j + e_j$, where the sampling error e_j is drawn from a mean-zero multivariate normal distribution with variance given by the empirical sampling covariance matrix of estimates for school j, denoted V_j.

The simulated oracle observes the true β_j's, while simulated parents observe some elements of $\hat{\beta}_j$ along with V_j and form forecasts of β_j^K to solve problem (10). Parents treat the population distribution of value-added as a prior and update beliefs based on information $\mathcal{P}$ according to Bayes' rule. For example, the posterior mean for college value-added (outcome

K) given a test score value-added estimate (outcome 1) is

$$E[\beta_j^K|\hat{\beta}_j^1, V_j^1] = \mu^K + \left(\frac{\sigma_{1K}}{\sigma_{11} + V_j^1}\right)\left(\hat{\beta}_j^1 - \mu^1\right), \tag{12}$$

where V_j^1 is the sampling variance (squared standard error) of $\hat{\beta}_j^1$. When the information portfolio $\mathcal{P}$ consists of estimates and standard errors for outcome 1, a parent orders schools by the predicted value in equation (12) and selects the school with the highest prediction.

To generate a realistic distribution of value-added facing each family we limit the choice set $\mathcal{C}$ to schools in a family's home borough. The analysis is conducted separately by choice set, including calculation of the bias-corrected covariance estimates that determine the prior. We then average simulation results across boroughs, weighting by the number of students.

Regret calculations quantify the extent to which survey information produces more effective forecasts of effects on Regents diplomas. Estimated regret for decisions targeting Regents diploma receipt, reported in the first column of Table 5, indicates that a parent with no information (who therefore ranks schools randomly) experiences regret of 19 percentage points relative to an oracle choosing schools in the same borough. Including survey information in the information set reduces regret 39% compared to this random-choice benchmark, while test score value-added produces a regret reduction of around 22%.

Consistent with the earlier findings, survey levels provide more informative forecasts only for Regents diplomas. In contrast, test score value-added dominates survey information for advanced diplomas and for post-secondary outcomes. An uninformed parent misses out on 17 percentage points of college enrollment value-added and 11 percentage points of two-year college persistence value-added, as shown in columns 3 and 4. Forecasts based on surveys alone offer little regret reduction for these outcomes, but using test score value-added reduces regret by 20 to 32 percent. Moreover, adding survey levels to a parent's information set provides essentially no reduction in regret once the parent knows test score VAM.

Test information based on SAT rather than Regents math scores proves even more valuable for parents targeting college attainment. Appendix Table A13 shows that school choice decisions guided by SAT VAM estimates reduce regret by around 36 percent for both college enrollment and persistence. As with the regression-based comparisons of alternative test VAMs described in Section 4.3, SAT VAM especially improves forecasts of school effects on college persistence. We again find minimal gains in advanced diploma and college outcomes from adding survey results once the information portfolio $\mathcal{P}$ includes SAT value-added.

Test score information also proves more valuable than survey information in choice environments with richer heterogeneity. Appendix Table A14 reports regret separately for Black or Hispanic students, female students, and students qualifying for free or reduced price lunch, which use college and test VAMs and survey levels estimated by student subsamples (the corresponding regression estimates appear in Appendix Table A15). These calculations extend the constant school effects model described in equation (1) to accommodate heterogeneity by demographics for college, test score, and survey effects. Regret simulations also tailor the parent's information portfolio $\mathcal{P}$ to demographic-specific estimates, which New York's school reports detail for some surveys and test score measures. Comparisons of regret with test and survey information in this extended choice setting present an arguably stronger case for test score over survey information.

The bottom of Table 5 reports effects of giving parents estimates of longer-term value-added (high school graduation or college attendance) rather than short-term estimates based on test scores and surveys. Here, regret relative to the oracle emerges only due to sampling error in the long-term VAM estimate $\hat{\beta}_j^K$. Since the standard errors of our estimates are modest, providing estimates of longer-term value-added improves decisions substantially, cutting regret by 86 to 97 percent across outcomes. The fact that direct estimates of college effects are better predictors of college value-added than indirect forecasts based on tests and surveys should not be surprising. At the same time, the large improvements in regret generated by direct access to high school graduation and college VAM highlight the low overall explanatory power of test scores and surveys for longer-term value-added, a finding that is also reflected in the low r-squared values in Tables 3 and 4. While test scores provide better longer-term forecasts than surveys for advanced diplomas and college, the contributions of schools to longer-term outcomes evidently operate through channels that aren't fully captured by either of these short-term measures.

6 Summary and Conclusions

US public school districts increasingly assess school quality based in part on surveys of school climate and student satisfaction. Many districts, including New York, have come to emphasize surveys over achievement tests in their accountability schemes. This paper investigates the value of school surveys for prediction of school causal impacts on consequential high school graduation and college attendance outcomes. Survey information is compared with the predictive value of test score value-added for these outcomes. Our analysis shows that survey data generates better forecasts of school effects on Regents diplomas than does test score data. But test value-added predicts impacts on longer-term outcomes more reliably than do school surveys. For advanced diploma and college outcomes, surveys provide little incremental value beyond test scores. From the point of view of parents seeking to boost their children's odds of going to college, test information is most valuable.

Although test VAM beats survey measures in forecasting impacts on several consequential outcomes, test scores may miss aspects of school quality important to parents, such as safety and socioemotional support, captured by surveys. It's important to keep in mind, however, that most variation in graduation and college value-added is unexplained by either tests or surveys. In the 2023-2024 school year, the New York City Public Schools added measures of high school graduation and college value-added to its School Performance Dashboard.[14] The effect of this new information on school choice and student outcomes is an important topic for future research.

References

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.

——— (2024a): "Credible School Value-Added with Undersubscribed School Lotteries," *Review of Economics and Statistics*, 106, 1–19.

——— (2024b): "Race and the Mismeasure of School Quality," *American Economic Review: Insights*, 6, 20–37.

Beuermann, D. W., C. K. Jackson, L. Navarro-Sola, and F. Pardo (2022): "What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output," *The Review of Economic Studies*, 90, 65–101.

Braga, M., M. Paccagnella, and M. Pellizzari (2014): "Evaluating Students' Evaluations of Professors," *Economics of Education Review*, 41, 71–88.

Bryk, A., P. B. Sebring, E. Allensworth, S. Luppescu, and J. Q. Easton (2010): *Organizing Schools for Improvement: Lessons from Chicago*, University of Chicago Press.

Carrell, S. E. and J. E. West (2010): "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy*, 118, 409–432.

Corradini, V. (2024): "Information and Access in School Choice Systems: Evidence from New York City," Working paper, MIT Economics.

[14] These measures were constructed by MIT's Blueprint Labs using methods similar to those in this paper.

Cunha, F. and J. Heckman (2007): "The Technology of Skill Formation," *American Economic Review*, 97(2), 31–47.

Deming, D. J. (2023): "Chapter 6 - Multidimensional Human Capital and the Wage Structure," Elsevier, vol. 7 of *Handbook of the Economics of Education*, 469–504.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly (2007): "Grit: Perservance and Passion for Long-Term Goals," .

Dweck, C. (2006): "Mindset: The New Psychology of Success," Random House.

Hart, H., C. Young, A. Chen, A. Zou, and E. M. Allensworth (2020): *Supporting School Improvement: Early Findings from Reexamination of the 5Essentials Survey*, University of Chicago Consortium on School Research.

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 26, 411–482.

Hernández, J. C. (2013): "Bloomberg Issues Final Letter Grades for New York Schools," Chalkbeat.

Hoffmann, F. and P. Oreopoulos (2009): "Professor Qualities and Student Achievement," *The Review of Economics and Statistics*, 91, 83–92.

Jackson, C. K. (2018): "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes," *Journal of Political Economy*, 126, 2072 – 2107.

Jackson, K. C., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020): "School Effects on Socioemotional Development, School-Based Arrests, and Educational Attainment," *American Economic Review: Insights*, 2, 491–508.

Jordan, P. W. and L. S. Hamilton (2020): "Walking a Fine Line: School Climate Surveys in State ESSA Plans," FutureEd. https://www.future-ed.org/wp-content/uploads/2020/01/FutureEdSchoolClimateReport.pdf.

Learning Policy Institute (2024): "Using Data to Improve School Climate," https://learningpolicyinstitute.org/media/4488/download?inline&file=Data_School_Climate_FACTSHEET.pdf, factsheet.

Merrill, L., C. Lafayette, and S. Goldberg (2018): "Redesigning the Annual NYC School Survey: Lessons from a Research-Practice Partnership," Research Alliance for NYC Schools, NYU Steinhardt, Available at: https://steinhardt.nyu.edu/sites/default/files/2021-01/Lessons_from_a_Research-Practice_Partnership.pdf.

Panorama Education (2025): https://www.panoramaed.com/about.

Rose, E. K., J. T. Schellenberg, and Y. Shem-Tov (2022): "The Effects of Teacher Quality on Adult Criminal Justice Contact," NBER working paper no. 30274.

Schneider, J., J. Noonan, R. S. White, D. Gagnon, and A. Carey (2021): "Adding "Student Voice" to the Mix: Perception Surveys and State Accountability Systems," *AERA Open*, 7.

Stark, P. B. and R. Freishtatd (2014): "An Evaluation of Course Evaluations," *ScienceOpen Research*.

Taylor, K. (2014): "New School Evaluations Will Lower Test Scores' Influence," *The New York Times*.

Veiga, C. (2017): "New York City Offers SAT to All High

School Juniors, Hoping to Clear a Path to College," *Chalkbeat*, https://www.chalkbeat.org/newyork/2017/4/3/21099704/new-york-city-offers-sat-to-all-high-school-juniors-hoping-to-clear-a-path-to-college

Wall, P. (2014): "Under Gentler Rating System, Schools Will No Longer Be Ranked or Graded," *Chalkbeat*, https://www.chalkbeat.org/newyork/2014/10/1/21092055/under-gentler-rating-system-schools-will-no-longer-be-ranked-or-graded/.

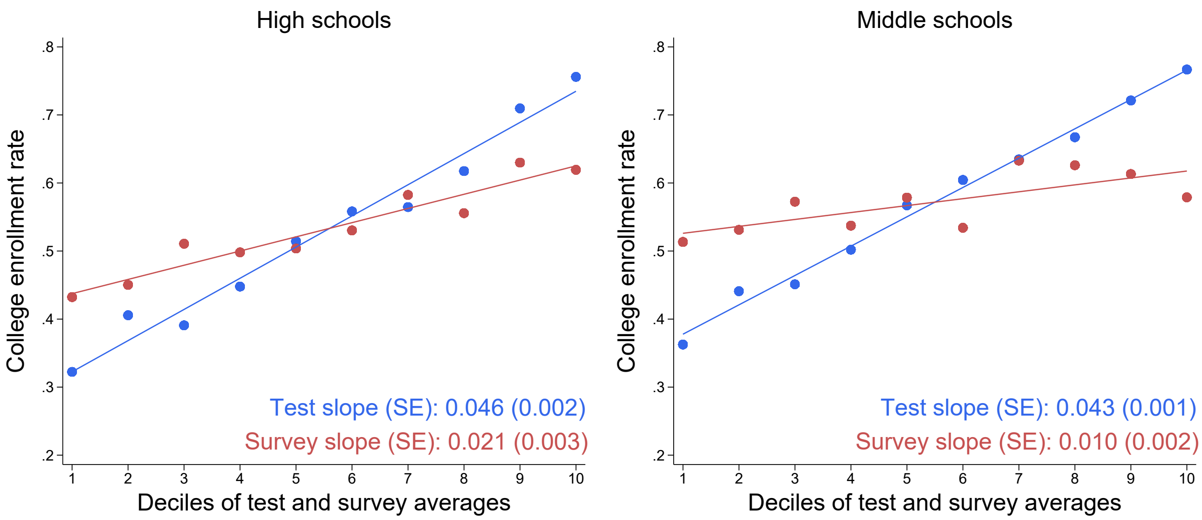
Walters, C. (2024): "Chapter 3 - Empirical Bayes Methods in Labor Economics," Elsevier, vol. 5 of *Handbook of Labor Economics*, 183–260.

Weinberg, B. A., M. Hashimoto, and B. M. Fleisher (2010): "Evaluating Teaching in Higher Education," *The Journal of Economic Education*, 23, 227–261.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

Zimmerman, A. (2024): "Top NYC Students Get Automatic SUNY Admission, but Fine Print ExcludesMany Black and Latino Kids," *Chalkbeat New York*, https://www.chalkbeat.org/newyork/2024/12/10/suny-admissions-promise-excludes-many-black-latino-kids-in-nyc/.

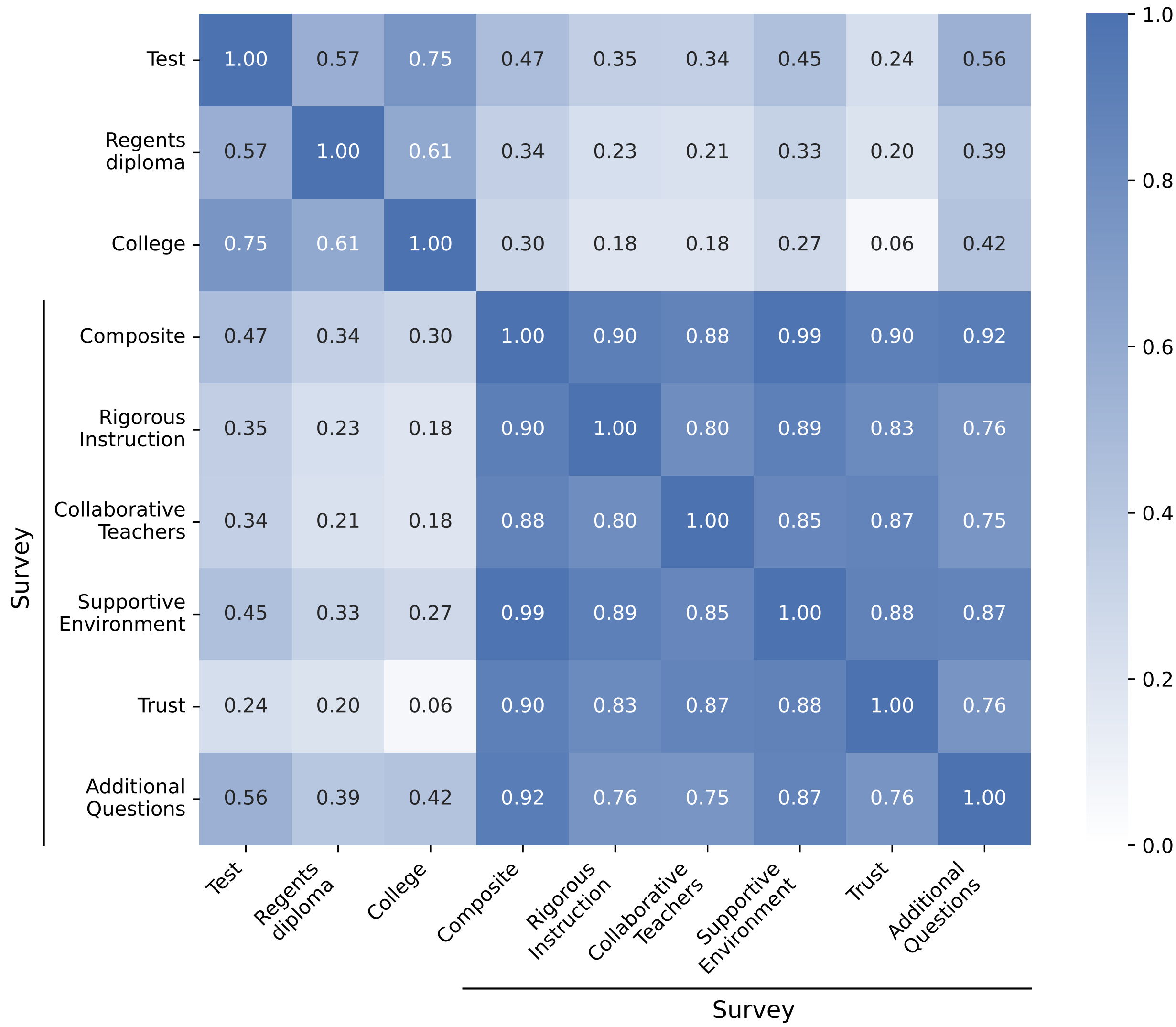
Figure 1. College Enrollment by Test and Survey School Averages



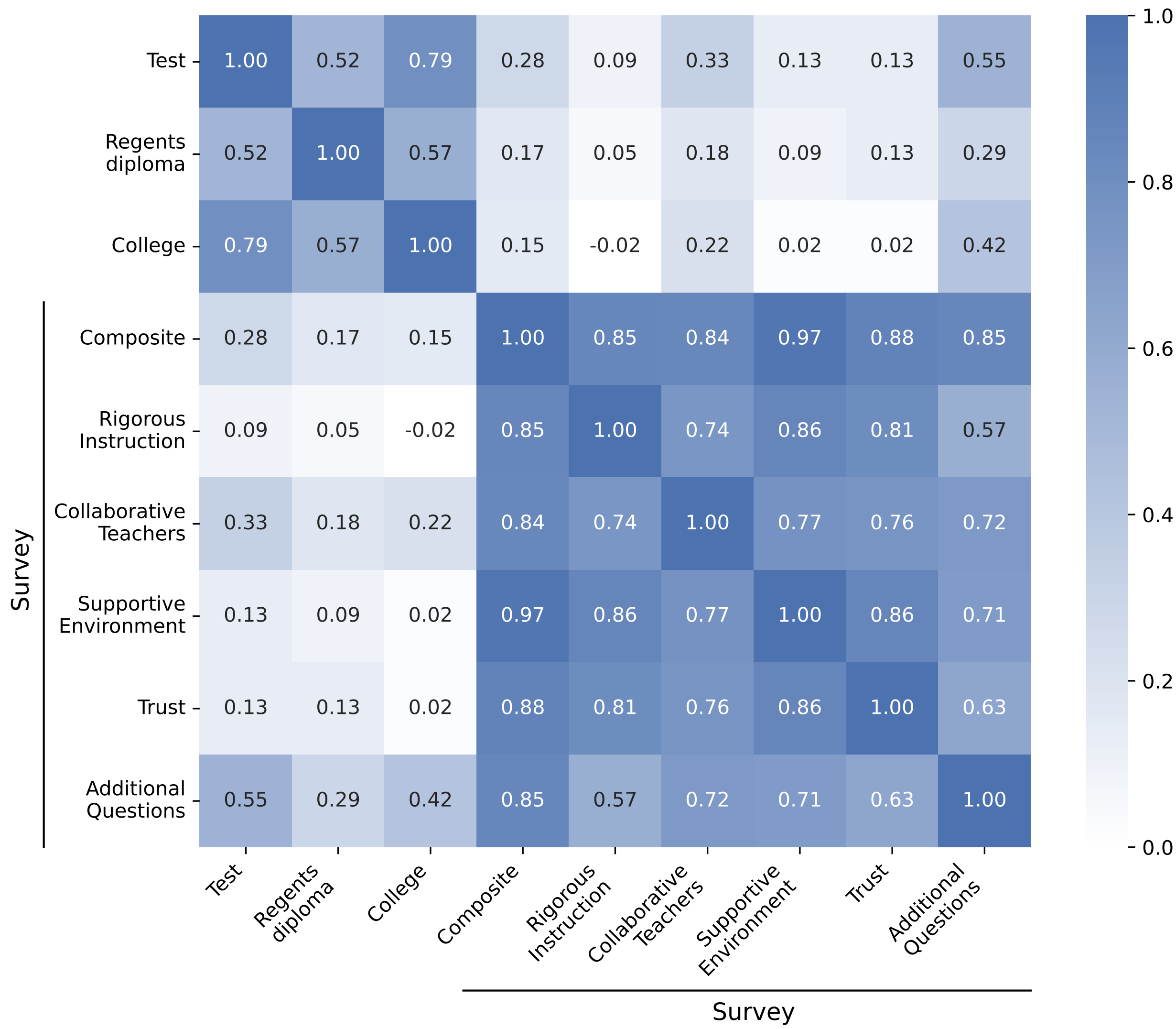
Notes: This figure plots average college enrollment rates against deciles of test scores and survey averages. Each point is a decile of schools by average test scores or average survey responses. The vertical axis measures the average college enrollment rate for schools in each decile. Slopes come from regressions of school college enrollment rates on test or survey decile indicators. College and test scores are measured for incoming high schoolers in fall 2012-2016, surveys for incoming high schoolers in fall 2014-2016 (due to survey redesign in 2014), and college, test scores, and survey outcomes for incoming middle schoolers in fall 2016. The test is Regents math in high school and the New York State sixth-grade state assessment in middle school. The survey is an index constructed from the NYC School Survey; see the Data Appendix for details.

Figure 2. Correlations between School Achievement Levels and Surveys

A. High schools



B. Middle schools



Notes: These figures display school-level correlations between measures of average student achievement and survey responses. High school test scores come from Regents Algebra 1 and Geometry exams taken by first-time 9th graders in school years 2012-2013 through 2016-2017. Middle school test scores come from state 6th-grade math and ELA exams. Regents diploma receipt and college enrollment are measured 4 years after 9th grade entry. The survey composite is an index constructed from the NYC School Survey for high school students enrolled in 9th grade in the 2014-2015 through 2016-2017 school years and for middle school students enrolled in 6th grade in the 2015-2016 school year. Other survey measures are constructed similarly for questions in categories defined by the NYC Framework for Great Schools. See the Data Appendix for details.

Table 1. Samples

	High school		Middle school	
	All (1)	With risk (2)	All (3)	With risk (4)
Demographics				
Hispanic	0.433	0.467	0.403	0.406
Black	0.289	0.275	0.271	0.224
Asian	0.143	0.146	0.164	0.199
White	0.126	0.104	0.148	0.159
Female	0.494	0.504	0.494	0.486
Free/reduced price lunch	0.786	0.794	0.738	0.733
Special education	0.189	0.072	0.207	0.191
English language learner	0.114	0.099	0.112	0.114
Baselines (std.)				
Math	-0.090	-0.026	0.032	0.100
ELA	-0.071	0.008	0.016	0.065
Enrollment				
Lottery tie-breaking	0.869	0.926	0.952	0.973
Screened (non-lottery tie-breaking)	0.131	0.074	0.048	0.027
Share enrolled where offered	0.660	0.706	0.634	0.635
On-time graduation and college				
Regents diploma	0.717	0.759	0.838	0.854
Advanced diploma	0.212	0.204	0.391	0.438
College enrollment	0.580	0.608	0.615	0.633
College persistence	0.365	0.371		
Share not offered	0.222	0.106	0.213	0.150
Students	269,492	77,057	68,528	24,612
Schools	483	483	616	584
Lotteries (schools with risk)		429		406

Notes: This table describes the middle and high school student samples used to estimate school value-added. Column 1 describes high school students enrolled in 9th grade in the 2012-2013 through 2016-2017 school years. Column 3 describes middle school students enrolled in 6th grade in the 2015-2016 school year. Columns 2 and 4 describe corresponding samples of applicants with assignment risk for at least one school. Baseline characteristics and lagged scores are from 5th grade for middle school students and 7th and 8th grade for high school students. 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 enrolling where offered does not include students receiving no offer. Graduation and college outcomes are detailed in the Data Appendix. College persistence data are not available for the middle school cohort.

Table 2. Tests For Bias in Value-Added Models

	Short-term outcomes				Longer-term outcomes			
	Test scores		Survey		Regents diploma		College enrollment	
	Levels	VAM	Levels	VAM	Levels	VAM	Levels	VAM
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: High schools								
Forecast coefficient	0.416	1.02	0.950	0.990	0.132	0.796	0.296	0.936
	(0.019)	(0.029)	(0.036)	(0.036)	(0.028)	(0.047)	(0.029)	(0.051)
Bias tests								
Forecast bias	923	0.719	1.97	0.070	956	18.7	591	1.59
	[0.000]	[0.396]	[0.160]	[0.791]	[0.000]	[0.000]	[0.000]	[0.208]
Omnibus (20 d.f.)	1029	35.9	25.9	15.4	1010	51.7	740	27.2
	[0.000]	[0.016]	[0.170]	[0.756]	[0.000]	[0.000]	[0.000]	[0.129]
First-stage F	1229	1308	854	889	1644	1507	1690	1547
Risk sample	64336	65864	36143	36340	63556	63866	65369	66958
Lottery schools	425	425	424	424	429	429	429	429
Panel B: Middle schools								
Forecast coefficient	0.294	1.06	0.956	1.00	0.510	0.878	0.352	0.820
	(0.043)	(0.065)	(0.058)	(0.059)	(0.129)	(0.120)	(0.133)	(0.144)
Bias tests								
Forecast bias	272	0.762	0.565	0.004	14.4	1.03	23.7	1.56
	[0.000]	[0.383]	[0.452]	[0.948]	[0.000]	[0.310]	[0.000]	[0.211]
Omnibus (20 d.f.)	314	22.6	20.8	17.4	31.5	20.2	38.2	30.5
	[0.000]	[0.308]	[0.409]	[0.624]	[0.049]	[0.446]	[0.008]	[0.062]
First-stage F	66.6	159	158	148	99.4	128	84.5	122
Risk sample	22875	23415	22337	22370	18119	18068	18738	18821
Lottery schools	405	406	403	403	404	406	405	406

Notes: This table reports tests for selection bias in school outcome levels and estimates from value-added models (VAMs). VAM estimates come from OLS regressions controlling for cubic functions of baseline math and ELA scores and indicators for sex, race, subsidized lunch, special education, and limited English proficiency, each interacted with application year. VAMs also include assignment propensity scores and running variable controls. High school VAMs add cubic functions of baseline attendance, disciplinary incidents, GPA, and credit accumulation. Forecast coefficients come from instrumental variables regressions of student outcomes on VAM fitted values, instrumenting fitted values with binned assignment indicators. Assignments are binned by ventile of the estimated VAM. IV models control for propensity scores, running variable controls, and baseline demographics and achievement. Test and survey outcomes are standardized to be mean zero and standard deviation one in the student-level distribution, separately by grade and year. The forecast bias test checks whether the forecast coefficient equals 1. The omnibus test combines the forecast test with an overidentification test that checks overidentifying restrictions implicit in the procedure used to estimate the forecast coefficient. Standard errors are reported in parentheses; test p-values are reported in brackets.

Table 3. Predicting School Effects on High School Graduation

	Regents diploma			Advanced diploma		
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: High schools			
Regressors						
Test score VAM	0.091		0.033	0.111		0.122
	(0.020)		(0.021)	(0.020)		(0.022)
Survey levels		0.106	0.096		0.016	-0.020
		(0.013)	(0.014)		(0.013)	(0.014)
R-squared	0.065	0.180	0.187	0.081	0.004	0.086
Schools		452			452	
Outcome mean		0.717			0.212	
			Panel B: Middle schools			
Regressors						
Test score VAM	0.019		-0.002	0.080		0.063
	(0.029)		(0.030)	(0.027)		(0.029)
Survey levels		0.026	0.026		0.031	0.022
		(0.011)	(0.012)		(0.011)	(0.012)
R-squared	0.002	0.019	0.019	0.032	0.025	0.043
Schools		591			591	
Outcome mean		0.838			0.391	

Notes: This table reports regressions of high school graduation value-added on test value-added and survey levels. Regression coefficients and r-squared's are derived from the estimated VAM covariance matrix and are bias-corrected as detailed in Section 3.2. Standard errors on reported coefficients are derived from the delta method and are also bias-corrected. Diploma outcomes are measured within 4 years of high school entry. See Data Appendix for details.

Table 4. Predicting School Effects on College Enrollment and Persistence

	College enrollment			College persistence		
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: High schools			
Regressors						
Test score VAM	0.107		0.089	0.046		0.040
	(0.019)		(0.021)	(0.013)		(0.015)
Survey levels		0.058	0.031		0.022	0.010
		(0.013)	(0.014)		(0.009)	(0.010)
R-squared	0.105	0.060	0.119	0.051	0.024	0.055
Schools		466			466	
Outcome mean		0.580			0.365	
			Panel B: Middle schools			
Regressors						
Test score VAM	0.165		0.179			
	(0.025)		(0.027)			
Survey levels		0.009	-0.019			
		(0.012)	(0.012)			
R-squared	0.146	0.002	0.154			
Schools		594				
Outcome mean		0.615				

Notes: This table reports regressions of college value-added on test value-added and survey levels. Regression coefficients and r-squared's are derived from the estimated VAM covariance matrix and are bias-corrected as detailed in Section 3.2. Standard errors on reported coefficients are derived from the delta method and are also bias-corrected. College enrollment is defined as enrollment in a 2- or 4-year college within 6 months of on-time high school graduation. College persistence indicates consecutive term enrollments for or graduation within 5 semesters after on-time enrollment. College persistence data are available only for the high school sample. See Data Appendix for details.

Table 5. Regret Relative to Oracle School Choice

	Regents diploma	Advanced diploma	College enrollment	College persistence
	(1)	(2)	(3)	(4)
Applicant's information set:				
Nothing (random choice)	0.189	0.211	0.166	0.112
Regret avoided	[-0%]	[-0%]	[-0%]	[-0%]
Test score VAM	0.146	0.151	0.113	0.090
	[-22.4%]	[-28.3%]	[-32.1%]	[-19.5%]
Survey levels	0.115	0.198	0.134	0.100
	[-38.8%]	[-6.4%]	[-19.5%]	[-10.0%]
Test score VAM, survey levels	0.114	0.151	0.111	0.090
	[-39.7%]	[-28.7%]	[-33.0%]	[-19.7%]
Longer-term VAM	0.013	0.005	0.016	0.016
	[-93.1%]	[-97.4%]	[-90.2%]	[-85.5%]
Oracle value-added	0	0	0	0
	[-100%]	[-100%]	[-100%]	[-100%]

Notes: This table reports average regret relative to oracle school choice for high school match applicants who seek to maximize value-added on the outcome listed in the column header. An applicant's regret is the maximum value-added of a school in their choice set net the value-added of their chosen school, averaged over realizations of value-added and information. An applicant selects the school in their borough of residence with the highest posterior mean value-added given a portfolio of information containing estimated test score VAM, survey averages, or longer-term VAM. Applicants are assumed to know the population distribution of VAM across schools and the sampling variances of VAM estimates. Longer-term VAM differs from oracle value-added in that the former has estimation error. Applicants without information pick a random school. Regret is computed via simulations calibrated to match VAM and levels covariance estimates calculated separately by borough. The change in regret relative to random choice is reported in square brackets.

Putting School Surveys to the Test: Supplemental Appendix

Joshua Angrist, MIT & NBER
Peter Hull, Brown University & NBER
Russell Legate-Yang, MIT
Parag Pathak, MIT & NBER
Christopher Walters, UC Berkeley & NBER

March 2025

A Appendix Figures and Tables

Table A1.A. Tests for Statistical Balance for High Schools

	Test scores		Survey		Regents diploma		College enrollment	
	Uncontrolled	Controlled	Uncontrolled	Controlled	Uncontrolled	Controlled	Uncontrolled	Controlled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Demographics								
Hispanic	-0.015	-0.004	0.012	0.002	0.058	-0.001	0.009	-0.007
	(0.001)	(0.004)	(0.001)	(0.003)	(0.001)	(0.004)	(0.001)	(0.003)
Black	-0.084	0.000	-0.075	-0.004	-0.016	-0.007	-0.022	0.002
	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)
Asian	0.070	-0.001	0.046	-0.000	-0.004	0.005	0.022	0.003
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
White	0.028	0.004	0.016	0.003	-0.040	0.003	-0.010	0.002
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
Female	0.009	0.003	0.016	0.000	-0.016	-0.002	0.035	-0.002
	(0.001)	(0.004)	(0.001)	(0.003)	(0.001)	(0.004)	(0.001)	(0.003)
Free/reduced price lunch	-0.025	-0.005	-0.028	-0.003	0.012	-0.003	-0.018	-0.004
	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)
Special education	-0.015	0.001	-0.013	0.004	-0.002	0.001	-0.002	-0.002
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
English language learner	0.004	0.001	0.013	-0.000	0.022	-0.002	-0.004	-0.003
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)	(0.002)
Baselines								
Math	0.214	-0.006	0.185	-0.003	-0.034	0.013	0.101	0.012
	(0.002)	(0.005)	(0.002)	(0.005)	(0.003)	(0.006)	(0.002)	(0.005)
ELA	0.163	0.003	0.143	-0.004	-0.043	0.001	0.096	0.012
	(0.002)	(0.005)	(0.002)	(0.005)	(0.003)	(0.006)	(0.002)	(0.005)
Joint F test p-value	0.000	0.360	0.000	0.411	0.000	0.131	0.000	0.150
Students	269,546	77,066	269,546	77,066	269,546	77,066	269,546	77,066

Notes: This table reports balance statistics, estimated by regressing baseline covariates on the VAM of the offered school and an indicator for any offer. Rows report the estimated coefficient on offered VAM. Columns 2, 4, 6, and 8 control for expected VAM, any offer risk, and running variable controls. Robust standard errors are reported in parentheses. The joint F-test row reports p-values from the test of no imbalance among all baseline covariates.

Table A1.B. Tests for Statistical Balance for Middle Schools

	Test scores		Survey		Regents diploma		College enrollment	
	Uncontrolled	Controlled	Uncontrolled	Controlled	Uncontrolled	Controlled	Uncontrolled	Controlled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Demographics								
Hispanic	-0.033	0.006	-0.015	0.011	-0.047	-0.014	-0.022	-0.006
	(0.003)	(0.008)	(0.003)	(0.007)	(0.003)	(0.011)	(0.003)	(0.010)
Black	-0.064	-0.010	-0.042	-0.014	0.009	0.000	-0.038	-0.004
	(0.002)	(0.008)	(0.003)	(0.007)	(0.003)	(0.010)	(0.003)	(0.009)
Asian	0.040	0.008	0.006	-0.000	0.003	0.012	0.028	0.010
	(0.002)	(0.005)	(0.002)	(0.005)	(0.002)	(0.006)	(0.002)	(0.006)
White	0.056	-0.007	0.050	0.001	0.032	-0.002	0.029	-0.001
	(0.002)	(0.005)	(0.002)	(0.004)	(0.003)	(0.006)	(0.002)	(0.005)
Female	0.007	0.001	0.009	-0.012	-0.009	-0.021	0.012	-0.008
	(0.003)	(0.008)	(0.003)	(0.007)	(0.003)	(0.010)	(0.003)	(0.009)
Free/reduced price lunch	-0.047	-0.009	-0.040	-0.012	-0.054	0.009	-0.059	-0.004
	(0.002)	(0.007)	(0.002)	(0.006)	(0.003)	(0.008)	(0.003)	(0.008)
Special education	-0.020	-0.005	-0.013	0.003	-0.020	0.001	-0.021	0.002
	(0.002)	(0.007)	(0.002)	(0.006)	(0.003)	(0.010)	(0.002)	(0.008)
English language learner	-0.008	-0.006	-0.013	0.006	-0.022	-0.011	-0.005	0.001
	(0.002)	(0.005)	(0.002)	(0.005)	(0.002)	(0.007)	(0.002)	(0.006)
Baselines (std.)								
Math	0.175	0.032	0.135	0.005	0.103	0.019	0.144	0.026
	(0.005)	(0.016)	(0.005)	(0.014)	(0.007)	(0.021)	(0.005)	(0.018)
ELA	0.148	0.017	0.124	-0.004	0.119	0.015	0.138	0.023
	(0.005)	(0.016)	(0.005)	(0.015)	(0.007)	(0.021)	(0.005)	(0.019)
Joint F test p-value	0.000	0.142	0.000	0.233	0.000	0.089	0.000	0.729
Students	68,550	24,615	68,550	24,615	68,550	24,615	68,550	24,615

Notes: This table reports balance statistics, estimated by regressing baseline covariates on the VAM of the offered school and an indicator for any offer. Rows report the estimated coefficient on offered VAM. Columns 2, 4, 6, and 8 control for expected VAM, any offer risk, and running variable controls. Robust standard errors are reported in parentheses. The joint F-test row reports p-values from the test of no imbalance among all baseline covariates.

Table A2. Tests for Differential Attrition

	High school		Middle school	
	College	Graduation	College	Graduation
	(1)	(2)	(3)	(4)
Offered VAM				
College enrollment	0.001	-0.001	-0.004	-0.005
	(0.002)	(0.003)	(0.008)	(0.008)
Regents diploma	0.003	0.003	0.017	0.013
	(0.003)	(0.003)	(0.009)	(0.009)
Test scores	0.003	0.007	-0.001	-0.001
	(0.003)	(0.003)	(0.007)	(0.007)
Survey	0.008	0.013	0.002	0.000
	(0.002)	(0.002)	(0.006)	(0.006)
Mean follow-up rate	0.812	0.767	0.768	0.713
N	335,823		71,815	

Notes: This table reports tests for differential attrition, estimated by regressing an indicator for non-missing follow-up outcome on the VAM of the offered school and an indicator for any offer. All models control for expected VAM, any offer risk, and running variable controls. Robust standard errors are reported in parentheses.

Table A3.A. VAM Bias Tests for Additional Outcomes for High Schools

	Short-term outcomes		Longer-term outcomes					
	SAT	Survey (grade 11)	Advanced diploma	Regents diploma within 6 years	4-year college enrollment	College enrollment within 6 years	College persistence	4-year college persistence
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Forecast coefficient	1.01	0.966	1.28	0.524	1.16	0.894	0.871	1.15
	(0.056)	(0.039)	(0.037)	(0.050)	(0.060)	(0.054)	(0.078)	(0.085)
Bias tests								
Forecast bias	0.026	0.755	58.8	89.0	6.95	3.91	2.76	2.95
	[0.871]	[0.385]	[0.000]	[0.000]	[0.008]	[0.048]	[0.097]	[0.086]
Omnibus (20 d.f.)	22.9	20.6	75.0	130	23.8	22.3	20.7	29.8
	[0.294]	[0.422]	[0.000]	[0.000]	[0.252]	[0.324]	[0.000]	[0.073]
First-stage F	1023	1072	1912	1513	1255	1456	1550	1494
Risk sample	50415	48890	63940	63873	66077	67121	66485	66958
Lottery schools	429	429	429	429	429	429	429	429

Notes: This table reports tests for selection bias in VAMs (columns 1 and 3-8) and school outcome levels (column 2) for additional outcomes. These estimates follow the procedures described in the note to Table 2. SAT scores are the average of standardized math and ELA subject scores and come from exams taken in the 3rd year of high school (grade 11 for most students). Survey responses come from the 3rd year of high school and are aggregated as described in Section 2. Diploma and college enrollment outcomes are measured within 4 years of high school entry or within 6 years, as noted in column headers. See Data Appendix for details.

Table A3.B. VAM Bias Tests for Additional Outcomes for Middle Schools

	Advanced diploma	4-year college enrollment
	(1)	(2)
Forecast coefficient	0.894	0.891
	(0.120)	(0.157)
Bias tests		
Forecast bias	0.777	0.482
	[0.378]	[0.488]
Omnibus (20 d.f.)	18.2	14.9
	[0.576]	[0.784]
First-stage F	154	102
Risk sample	17833	19041
Lottery schools	405	406

Notes: This table reports tests for bias in VAMs for additional high school graduation and college outcomes. These estimates follow the procedures described in the note to Table 2. Diploma and college enrollment outcomes are measured within 4 years of high school entry. See Data Appendix for details.

Table A4. Selection Bias in Test Scores and Survey Responses

	Test score	Survey
	(1)	(2)
Outcome on baselines and school FE		
Baseline math	0.647	-0.008
	(0.005)	(0.005)
Baseline survey	0.016	0.417
	(0.003)	(0.005)
Adjusted r-squared from baseline on school FE		
Baseline math	0.256	
Baseline survey	0.026	

Notes: The top panel reports coefficients on baseline math scores and survey responses from regressions of the outcomes specified in the column headers on these baseline measures and school fixed effects. The bottom panel reports adjusted r-squared's from regressions of the baseline scores specified in the row labels on school fixed effects. Standard errors are reported in parentheses. The sample for this analysis includes first-time 9th graders in school years 2015-2016 and 2016-2017. Baseline survey responses are unavailable for earlier cohorts.

Table A5. Standard Deviations of Levels and Value-Added

	Short-term outcomes		Longer-term outcomes	
	Test scores	Survey	Regents diploma	College enrollment
	(1)	(2)	(3)	(4)
		Panel A: High schools		
Levels	0.456	0.321	0.172	0.169
VAM	0.226	0.287	0.080	0.070
		Panel B: Middle schools		
Levels	0.510	0.450	0.106	0.143
VAM	0.202	0.439	0.084	0.087

Notes: This table reports standard deviations of school outcome levels and value-added for the outcomes listed in the column headers. Standard deviations are derived from the estimated VAM covariance matrix and are bias-corrected as detailed in Section 3.2.

Table A6. Predicting School Effects on College Enrollment with SAT VAM

	Regents diploma			Advanced diploma			College enrollment			College persistence		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors												
SAT VAM	0.067		0.008	0.165		0.168	0.186		0.167	0.147		0.145
	(0.040)		(0.040)	(0.042)		(0.044)	(0.037)		(0.038)	(0.027)		(0.028)
Survey levels		0.096	0.096		0.012	-0.005		0.048	0.031		0.017	0.002
		(0.014)	(0.015)		(0.014)	(0.015)		(0.014)	(0.014)		(0.010)	(0.010)
R-squared	0.010	0.131	0.131	0.052	0.002	0.052	0.104	0.042	0.121	0.160	0.013	0.160
Schools		458			458			462			462	
Outcome mean		0.717			0.212			0.580			0.365	

Notes: This table reports regressions of high school graduation and college value-added on SAT value-added and survey levels. The survey variable uses student responses in the 3rd year of high school (grade 11 for most students) to match the timing of the SAT test. See the Data Appendix for details. The models and derivation procedure used to compute these estimates are described in the note to Tables 3 and 4.

Table A7. Predicting School Effects on College Enrollment with Alternative Survey Measures

		First principal component		Rigorous Instruction		Collaborative Teachers		Supportive Environment		Trust		Additional questions	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
							Panel A: High schools						
Regressors													
Test score VAM	0.107		0.091		0.097		0.097		0.087		0.100		0.088
	(0.019)		(0.021)		(0.021)		(0.020)		(0.021)		(0.020)		(0.021)
Survey levels		0.055	0.028	0.068	0.025	0.067	0.030	0.064	0.035	0.050	0.024	0.051	0.029
		(0.013)	(0.014)	(0.019)	(0.021)	(0.020)	(0.021)	(0.014)	(0.015)	(0.017)	(0.018)	(0.011)	(0.012)
R-squared	0.105	0.055	0.117	0.039	0.109	0.036	0.111	0.065	0.121	0.026	0.111	0.064	0.121
							Panel B: Middle schools						
Regressors													
Test score VAM	0.165		0.182		0.196		0.180		0.184		0.180		0.142
	(0.025)		(0.026)		(0.026)		(0.027)		(0.026)		(0.026)		(0.028)
Survey levels		0.003	-0.024	-0.032	-0.070	0.008	-0.032	-0.008	-0.034	-0.031	-0.050	0.045	0.021
		(0.012)	(0.012)	(0.019)	(0.019)	(0.019)	(0.019)	(0.013)	(0.013)	(0.016)	(0.015)	(0.011)	(0.012)
R-squared	0.146	0.000	0.159	0.012	0.201	0.001	0.157	0.001	0.169	0.015	0.183	0.070	0.158

Notes: This table reports regressions of college value-added on test value-added and survey levels, measured as described in the column header. Column 1 reproduces the estimates from Table 4. Columns 2 and 3 use the first principal component of standardized survey item responses. Columns 4-11 use the average of item responses to questions within the survey category in the column header. These categories are defined by New York's "Framework for Great Schools." Columns 12 and 13 report results for the average of uncategorized survey questions. See the Data Appendix for details on the construction of these survey measures. The models and derivation procedure used to compute these estimates are described in the note to Table 4.

Table A8. Predicting School Effects on 4-Year College Enrollment and Persistence

	4-year college enrollment			4-year college persistence		
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: High schools			
Regressors						
Test score VAM	0.094		0.082	0.042		0.036
	(0.017)		(0.019)	(0.012)		(0.013)
Survey levels		0.045	0.020		0.021	0.010
		(0.011)	(0.012)		(0.008)	(0.008)
R-squared	0.102	0.046	0.110	0.062	0.031	0.068
Schools		466			466	
Outcome mean		0.420			0.290	
			Panel B: Middle schools			
Regressors						
Test score VAM	0.160		0.169			
	(0.024)		(0.025)			
Survey levels		0.014	-0.012			
		(0.011)	(0.011)			
R-squared	0.173	0.006	0.178			
Schools		594				
Outcome mean		0.500				

Notes: This table reports regressions of 4-year college value-added on test value-added and survey levels. The models and derivation procedure used to compute the estimates are described in the note in Table 4. 4-year college enrollment is defined as enrollment in a 4-year college within 6 months of on-time high school graduation. 4-year college persistence indicates consecutive term enrollments for or graduation within 5 semesters after on-time enrollment at a 4-year college. College persistence data are available only for the high school sample. See Data Appendix for details.

Table A9. Predicting School Effects on High School Graduation and College Enrollment Within 6 Years

	Regents diploma within 6 years			College enrollment within 6 years		
	(1)	(2)	(3)	(4)	(5)	(6)
Regressors						
Test score VAM	0.108		0.049	0.101		0.084
	(0.017)		(0.017)	(0.017)		(0.019)
Survey levels		0.111	0.097		0.054	0.029
		(0.010)	(0.012)		(0.011)	(0.013)
R-squared	0.131	0.288	0.311	0.122	0.069	0.138
Schools		452			466	
Outcome mean		0.790			0.680	

Notes: This table reports regressions of college enrollment and Regents diploma value-added within 6 years of high school entry on test score value-added and survey levels. See Data Appendix for details on the construction of within 6 year graduation and college outcomes. These outcomes are available only for the high school sample. The models and derivation procedure used to compute these estimates are described in the note to Tables 3 and 4.

Table A10. IV VAM Predictions of School Effects on High School Graduation

	Regents diploma			Advanced diploma		
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: High schools			
Regressors						
Test score VAM	0.047		0.005	0.142		0.156
	(0.022)		(0.023)	(0.028)		(0.031)
Survey levels		0.071	0.069		0.023	-0.023
		(0.014)	(0.016)		(0.019)	(0.021)
Schools		424			424	
			Panel B: Middle schools			
Regressors						
Test score VAM	-0.008		-0.035	0.089		0.101
	(0.054)		(0.056)	(0.055)		(0.059)
Survey levels		0.029	0.035		0.001	-0.015
		(0.022)	(0.024)		(0.024)	(0.026)
Schools		423			423	

Notes: This table reports IV VAM estimates of regressions of graduation value-added on test value-added and survey levels. Regression coefficients are derived from 2SLS estimates of student graduation outcomes regressed on fitted values for graduation VAM, test VAM and survey levels, instrumented with individual school offers. These regressions control for assignment risk and VAM controls. 2SLS estimates are combined with the estimated covariance matrix between graduation VAM, test VAM, and survey levels to compute the "short" regression of graduation value-added on test VAM and/or survey levels via the omitted variables bias formula. The estimated VAM covariance matrix is bias-corrected as described in Section 3.2. Estimates of graduation VAM are not assumed to match graduation value-added parameters. Standard errors reflect uncertainty in the estimated 2SLS coefficients and the VAM covariance matrix; these are derived via the delta method. See Angrist et al. (2024a,b) for details on the IV VAM procedure.

Table A11. IV VAM Predictions of School Effects on College Enrollment and Persistence

	College enrollment			College persistence		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: High schools						
Regressors						
Test score VAM	0.083		0.076	0.048		0.045
	(0.026)		(0.029)	(0.022)		(0.024)
Survey levels		0.035	0.012		0.019	0.005
		(0.018)	(0.019)		(0.015)	(0.016)
Schools		424			424	
Panel B: Middle schools						
Regressors						
Test score VAM	0.114		0.127			
	(0.061)		(0.065)			
Survey levels		0.003	-0.016			
		(0.027)	(0.028)			
Schools		424				

Notes: Regression coefficients are derived from 2SLS estimates of student graduation outcomes regressed on fitted values for graduation VAM, test VAM and survey levels, instrumented with individual school offers. These regressions control for assignment risk and VAM controls. 2SLS estimates are combined with the estimated covariance matrix between graduation VAM, test VAM, and survey levels to compute the "short" regression of graduation value-added on test VAM and/or survey levels via the omitted variables bias formula. The estimated VAM covariance matrix is bias-corrected as described in Section 3.2. Estimates of graduation VAM are not assumed to match graduation value-added parameters. Standard errors reflect uncertainty in the estimated 2SLS coefficients and the VAM covariance matrix; these are derived via the delta method. See Angrist et al. (2024a,b) for details on the IV VAM procedure.

Table A12. Predicting School Effects on High School Graduation and College Enrollment with Conventional Test VAMs

	Regents diploma		College enrollment	
	(1)	(2)	(3)	(4)
	Panel A: High school			
Regressors				
Conventional test VAM	0.028		0.078	
	(0.021)		(0.020)	
Test progress		0.044		0.072
		(0.020)		(0.019)
Survey levels	0.098	0.096	0.034	0.041
	(0.014)	(0.013)	(0.014)	(0.013)
R-squared	0.185	0.195	0.105	0.105
Schools	458	458	462	462
	Panel B: Middle schools			
Regressors				
Conventional test VAM	-0.006		0.186	
	(0.020)		(0.021)	
Test progress		-0.017		0.169
		(0.021)		(0.021)
Survey levels	0.027	0.028	-0.018	-0.014
	(0.009)	(0.009)	(0.009)	(0.009)
R-squared	0.019	0.020	0.173	0.136
Schools	591	591	594	594

Notes: This table reports regressions of college value-added on conventional test VAM, test score progress, and survey levels. Conventional test score VAM controls for student demographics and lagged achievement but omits assignment risk controls. Test score progress measures are derived from conventional VAMs estimated with differences in test score outcomes from baseline; see Angrist et al. (2017) for details. The models and derivation procedure used to compute these estimates are described in the note to Tables 3 and 4.

Table A13. Regret Relative to Oracle School Choice with SAT VAM

	Regents diploma	Advanced diploma	College enrollment	College persistence
	(1)	(2)	(3)	(4)
Applicant's information set:				
Nothing (random choice)	0.190	0.212	0.155	0.108
Regret avoided	[-0%]	[-0%]	[-0%]	[-0%]
SAT VAM	0.176	0.165	0.101	0.068
	[-7.4%]	[-22.2%]	[-35.1%]	[-37.2%]
Survey levels	0.126	0.200	0.129	0.100
	[-33.8%]	[-5.8%]	[-16.5%]	[-7.0%]
SAT VAM, survey levels	0.126	0.165	0.098	0.068
	[-33.8%]	[-22.2%]	[-36.4%]	[-37.2%]
Longer-term VAM	0.013	0.005	0.017	0.016
	[-93.0%]	[-97.5%]	[-89.3%]	[-85.1%]
Oracle value-added	0	0	0	0
	[-100%]	[-100%]	[-100%]	[-100%]

Notes: This table reports average regret relative to oracle school choice for high school match applicants who seek to maximize value-added on the outcome listed in the column header. Applicants choose among schools in their borough of residence. The models and derivation procedure used to compute these estimates are described in the note to Tables 5.

Table A14. College Enrollment Regret Relative to Oracle School Choice by Subgroups

	Black or Hispanic	Female	Free/reduced price lunch
	(1)	(2)	(3)
Applicant's information set:			
Nothing (random choice)	0.174	0.168	0.163
Regret avoided	[-0%]	[-0%]	[-0%]
Test score VAM	0.113	0.127	0.105
	[-35.3%]	[-24.5%]	[-35.5%]
Survey levels	0.142	0.140	0.135
	[-18.1%]	[-17.0%]	[-17.5%]
Test score VAM, survey levels	0.112	0.125	0.105
	[-35.7%]	[-25.9%]	[-35.7%]
Longer-term VAM	0.019	0.027	0.019
	[-89.0%]	[-83.9%]	[-88.2%]
Oracle value-added	0	0	0
	[-100%]	[-100%]	[-100%]

Notes: This table reports average regret relative to oracle school choice for high school match applicants in the demographic group listed in column headers who seek to maximize college value-added. These results are calculated by repeating the VAM estimation and regret simulation procedures separately for each subsample of students by demographic group, following the procedures described in the notes to Table 5. Applicants choose among schools in their borough of residence.

Table A15. Predicting School Effects on College Enrollment by Subgroups

	Black or Hispanic			Female			Free/reduced price lunch		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					Panel A: High schools				
Regressors									
Test score VAM	0.112		0.091	0.090		0.076	0.115		0.096
	(0.019)		(0.021)	(0.021)		(0.024)	(0.019)		(0.022)
Survey levels		0.066	0.039		0.045	0.024		0.061	0.031
		(0.014)	(0.015)		(0.014)	(0.016)		(0.013)	(0.015)
R-squared	0.116	0.074	0.138	0.072	0.039	0.082	0.121	0.066	0.135
Schools		460			449			461	
					Panel B: Middle schools				
Regressors									
Test score VAM	0.192		0.200	0.182		0.202	0.172		0.180
	(0.028)		(0.030)	(0.040)		(0.043)	(0.029)		(0.030)
Survey levels		0.019	-0.012		0.001	-0.027		0.018	-0.011
		(0.015)	(0.015)		(0.017)	(0.017)		(0.013)	(0.014)
R-squared	0.152	0.007	0.155	0.085	0.000	0.094	0.139	0.007	0.141
Schools		576			560			579	

Notes: This table reports regressions of college value-added on test score value-added and survey levels, separately by subgroup. These coefficients are calculated by estimating college and test VAM and survey levels in the subsamples of students described in the column headers. The models and derivation procedure used to compute these estimates by subsample are described in the note to Table 4.

Table A16. Predicting Expert Reviews

	College enrollment			College persistence			Regents diploma			Advanced diploma		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
						Panel A: High schools						
Regressors												
VAM	0.264		-0.059	0.204		-0.133	0.343		0.118	0.026		-0.284
	(0.057)		(0.062)	(0.065)		(0.076)	(0.054)		(0.047)	(0.050)		(0.059)
Levels		0.536	0.569		0.480	0.560		0.625	0.579		0.350	0.519
		(0.045)	(0.057)		(0.045)	(0.065)		(0.043)	(0.046)		(0.047)	(0.059)
R-squared	0.069	0.286	0.288	0.041	0.229	0.240	0.117	0.389	0.401	0.001	0.122	0.173
						Panel B: Middle schools						
Regressors												
VAM	0.265		-0.221				0.130		-0.305	0.286		-0.284
	(0.072)		(0.092)				(0.074)		(0.083)	(0.059)		(0.113)
Levels		0.557	0.708					0.502	0.693		0.486	0.713
		(0.044)	(0.075)					(0.047)	(0.071)		(0.043)	(0.101)
R-squared	0.070	0.310	0.335				0.017	0.251	0.308	0.082	0.236	0.265

Notes: This table reports regressions of expert ratings on VAM and school outcome levels for the outcome specified in the column header. Expert ratings, VAM, and levels are standardized to have mean zero and variance one across schools. Regression coefficients and r-squared's are derived from bias-corrected VAM, levels, and expert rating covariance estimates, as described in Section 3.2. Expert ratings are assumed to have no sampling variance. Standard errors are derived via the delta method and are also bias-corrected. The samples in panels A and B include 438 and 498 schools, respectively.

B Evaluating Expert Reviews

Alongside surveys, New York's school performance framework features ratings by outside experts sent to visit and evaluate schools, called "quality reviews." Though found less often in other school districts than surveys, expert reviews and surveys gained similar prominence over test scores in New York's 2014 accountability redesign (Taylor, 2014; Wall, 2014). Expert reviewers, often former educators or principals, sit in on classes and meet with students, parents, and teachers to evaluate how well schools are organized to support student learning. Expert reviews and surveys are both organized around New York's "Framework for Great Schools," an intellectual program for school quality inspired by Bryk et al. (2010).

Given these similarities, we also evaluate expert reviews as school quality measures. In the VAM framework, expert reviews are not linked to or measured in units of a student outcome, so they can't be interpreted like test or survey VAM. College VAM predictions using test and survey causal effects seem likely to be informative in other settings because these correlations may reflect common patterns in how schools produce learning gains. Since New York's expert review program is somewhat unique, our analysis separates expert reviews from tests and surveys and evaluates how closely reviews track school causal effects on longer-term outcomes.[15] Alternatively, expert ratings based on student meetings and classroom observations may also reflect selection bias.

Estimates in Table A16 show that expert reviews are more related to selection bias than to school causal effects. This table reports regressions that predict a composite measure of a school's expert review with VAMs and levels for college enrollment, college persistence, Regents diplomas, and advanced diplomas, calculated as in Section 3.2. Since expert ratings have no natural unit in terms of student outcomes, variables are standardized to have variance one across schools and regression coefficients are partial correlation estimates. See the Data Appendix for details.

Column 1 of Panel A shows that a one standard deviation improvement in college VAM predicts a 0.26 standard deviation increase in expert rating for high schools, a modest positive relationship. The corresponding correlation with college levels in column 2 is more than twice as large. Regressions in column 3 that include both VAM and levels predictors find an unchanged levels coefficient and a negative though not statistically significant VAM coefficient.

The negative VAM coefficient in column 3 combined with VAM regression algebra demonstrates that experts prefer schools that achieve a given college-going rate by enrolling stu-

[15] Predictions using levels of college-going and graduation rates or average test scores also seem likely to reflect idiosyncratic patterns in school demand and access, residential sorting, and information rather than relationships between causal effects.

dents likely to attend college at baseline instead of improving college-going odds. Grouping the VAM regression in equation (4) by school under the selection-on-observables restriction $E[e_i^k|D_{ij}=1]=0$ yields

$$E[Y_i^k|D_{ij}=1]=\beta_j^k+E[X_i'|D_{ij}=1]\gamma^k. \tag{13}$$

Equation (13) decomposes college levels into causal effects and a school's average college-going odds according to controls X_i, a measure of selection bias. Regressing expert ratings on these two components generates a selection bias coefficient equal to the levels effect in column 3 and a VAM coefficient equal to the levels plus VAM effects. Given the negative VAM effect, schools that aim to improve expert reviews do better by admitting more positively-selected students than by improving outcomes for a fixed student population.[16]

Regressions of expert reviews on VAMs and levels for college persistence and advanced diploma receipt reveal arguably stronger selection bias patterns. Given Regents graduation rates, expert reviews of high schools increase slightly in Regents VAM—though graduation rates remain a much better predictor on their own. For all four longer-term outcomes, higher quality middle schools receive significantly worse expert reviews given levels. Taken together, these estimates suggest that expert reviews offer little beyond unadjusted graduation and college-going rates as stand-ins for causal effects. More broadly, our results for surveys and expert reviews suggest that test score value-added is a better predictor of longer-term school effectiveness than qualitative information reported by either students or experts.

[16] VAM regression mechanics discussed here apply to un-scaled VAM, levels, and selection bias regressors. While VAM and levels in Table A16 are scaled to have variance 1, only the sign of the VAM coefficient in column 3, not its scale, determines whether the selection bias coefficient is larger or smaller than the VAM coefficient in the auxiliary regression.

C Data Appendix

Data Sources

New York City Public Schools (NYCPS) Assignment Data

Annual records from the NYCPS school assignment system provide data on middle school applications for the 2015-2016 school year from 5th grade applicants, as well as high school applications from 2012-2013 through 2016-2017 from 8th grade applicants. These records contain student rank-order lists, program priorities, lottery tiebreaker values, and final assignments. The data exclude applications to charter schools and New York's nine specialized high schools.

Student Demographics and Enrollment

The NYCPS Office of School Performance and Accountability (OSPA) provides annual end-of-year records from 2012-2022. These records contain student grade level and enrolled school information, along with demographic characteristics, attendance records, special education status, subsidized lunch status, and limited English proficiency status.

Academic Performance Data

State Assessments New York State Assessment scores in math and English Language Arts (ELA) for grades 3-8 are provided by NYCPS OSPA.

Regents test scores Regents math exam scores for grade 9 students are provided by NYCPS OSPA.

SAT SAT data is provided by College Board via NYCPS and matched annually to NYCPS records using student name and date of birth.

High School Graduation NYCPS OSPA provides data on high school diploma receipt and graduation timing. Student outcomes are tracked starting from initial 9th grade entry year, with status recorded at three points: 4, 5, and 6 years after entering high school. At each point, the data document whether students are still enrolled (including current grade level), have graduated (including type of diploma earned), or have dropped out. For the middle school sample, only the 4-year (on-time) graduation outcomes are available.

Post-Secondary Outcomes

The National Student Clearinghouse (NSC) provides data on postsecondary term enrollments via NYCPS. NYCPS provides NSC with student name and date of birth for matches. NYCPS searches students in NSC records three consecutive years after high school graduation. NSC data include enrollment term and dates, college name, college type (2 or 4 year), enrollment status, and major and degrees if applicable.

New York City School Quality Measures

School Climate Surveys Data on responses to the New York City School Survey are provided by NYCPS OSPA from 2014 to 2020. Records include student, teacher, staff, and parent responses.

Quality Reviews Data on quality reviews for academic years 2012-2013 through 2016-2017 come from publicly available records hosted by NYC Open Data. These reviews involve experienced educators (experts) conducting two-day school visits to evaluate schools across 10 distinct quality indicators, organized into New York's Framework for Great Schools. Charter schools do not receive expert reviews.

Key Outcome Variables

Academic Achievement

For middle school test scores, we use an average of math and ELA state assessments taken by 6th graders during the 2015-2016 school year. For high school test scores, we use the scores from Regents Algebra 1 and Geometry math exams taken by first-time 9th graders in school years 2012-2013 through 2016-2017. We also consider high school SAT test scores taken in the 3rd year of high school, typically corresponding to grade 11. The SAT test score measure averages math and ELA subject scores. All subject scores are standardized to mean zero and standard deviation one within each grade and year. Averages of standardized subject scores for the middle and high school test score measures are again standardized to have mean zero and standard deviation one for each grade in each year among all test-takers.

High School Graduation

New York State offers three types of high school diplomas: standard Regents diplomas available to all students; advanced Regents diplomas requiring additional proficiency in math, science, and foreign languages; and local diplomas with lower examination requirements and limited eligibility. Almost all students who graduate receive a Regents diploma in our sample period. On-time graduation is defined as graduation within 4 years of high school entry. Other measures include graduation within 6 years of high school entry. Local diplomas are not included in our analysis.

College Outcomes

We examine two primary college outcomes. The first is college enrollment, measured as enrollment within 6 months of on-time high school graduation, with separate indicators for any college and 4-year college enrollment. The second is college persistence, defined as completing the second year (either enrolled in 5th consecutive semester or graduated), conditional on on-time college enrollment. We define persistence separately for any college and 4-year college enrollment. Students missing from NSC records but confirmed as high school graduates or dropouts are coded as not enrolled in college.

Surveys

We construct a standardized survey composite that averages standardized student item responses. To be comparable with the timing of the test score outcome, we use middle school students' responses in grade 6 and high school students' responses in grade 9. In constructing the survey variable for each year, we include all available survey questions except those explicitly designed for specific grade levels (such as questions targeted only at grades 6-8 or 9-12). Item responses are recorded on a categorical 1 - 4 scale, with 4 indicating most agreement with the question. We recode survey items so higher values indicate more favorable responses according to New York's survey reports, which record how many students respond favorably to a question. We standardize item responses by grade and year. Our primary survey measures averages these standardized item responses, then standardizes this average by grade-year, consistent with our test score standardization scheme. Missing item responses are dropped in the average.

We also construct alternative survey measures: a PCA-based measure, and separate measures for survey categories defined by NYC's Framework for Great Schools (Rigorous Instruction, Collaborative Teachers, Supportive Environment, and Trust), with remaining questions classified as Additional Questions. For the PCA measure, we use the first principal component of standardized item responses within each grade-year, imputing missing survey responses with the mean. For the survey category measures, we use the same aggregation approach as our main survey average, but include only questions corresponding to that specific category.

Quality Reviews

Quality ("expert") reviews are based on experienced educators' ratings of schools on five quality indicators (Curriculum, Pedagogy, Assessment, High Expectations, and Teacher Teams and Leadership Development) for which we have data for most schools in our sample. Our review measure averages standardized indicator ratings and then standardizes this average to be mean zero and standard deviation one among schools. For high schools with multiple reviews between the 2012-2013 and 2016-2017 academic years, we use the average of all available expert reviews during this period. Our middle school expert review measures use reviews from the 2015-2016 school year when available and the 2014-2015 school year when not.