

Who Benefits from Attending Effective High Schools?

C. Kirabo Jackson, *Northwestern University*

Sebastián Kiguel, *Northwestern University*

Shanette C. Porter, *Student Experience Research Network*

John Q. Easton, *UChicago Consortium on School Research*

We estimate the longer-run effects of attending an effective high school (one that improves a combination of test scores, survey measures of socioemotional development, and behaviors in ninth grade) for students who are more versus less educationally advantaged. All students benefit from attending effective schools, but the least advantaged students experience larger improvements in high school graduation, college going, and school-based arrests. Test score value-added understates the long-run importance of effective schools, particularly for less advantaged populations. Patterns suggest that this may, in part, reflect less advantaged students being relatively more responsive to non-test-score dimensions of school quality.

We thank the staff at Chicago Public Schools, particularly the Office of Social and Emotional Learning and the UChicago Consortium on School Research, for providing access to and information about the Chicago Public Schools data. This paper benefited from discussion with seminar participants at the UChicago Consortium on School Research, and data management was facilitated by its archivist, Todd Rosenkranz. We acknowledge funding for this research from the Bill and Melinda Gates Foundation. The content is solely the responsibility of the authors. Contact the corresponding

Submitted March 11, 2022; Accepted February 7, 2023; Electronically published May 15, 2024.

Journal of Labor Economics, volume 42, number 3, July 2024.

© 2024 The University of Chicago. All rights reserved. Published by The University of Chicago Press in association with The Society of Labor Economists and The National Opinion Research Center. <https://doi.org/10.1086/724568>

I. Introduction

Schools can have important effects on students' longer-term outcomes. For example, some charter schools increase college going (e.g., Angrist et al. 2016; Sass et al. 2016); some selective schools improve educational attainment, wages, and health (e.g., Jackson 2010; Beuermann and Jackson 2022); and winning a school-choice lottery may increase college going for girls and reduce arrests among certain boys (e.g., Deming 2011; Deming et al. 2014). However, questions remain about whether the benefits of attending better schools differ for better- or worse-prepared students. We explore whether "effective schools" (i.e., those that improve both test scores and socioemotional development [SED]) confer similar longer-run impacts on more and less educationally advantaged (i.e., likely to attain more years of education based on eighth-grade characteristics) students.

In principle, the least educationally advantaged students may benefit most from effective schools because they may have more room for improvement. Conversely, if "skills beget skills" (Cunha, Heckman, and Schennach 2010), effective schools, on average, may have small impacts on the least advantaged. The existing empirical evidence on this topic is mixed and (to overcome selection issues) has been focused on small numbers of oversubscribed charter schools or elite schools that rely on admission lotteries or admission tests.¹ Moreover, because these studies rely on comparisons among applicants to these special oversubscribed schools (who may differ from typical students, as in Hoxby and Murarka 2009), patterns in these studies may differ from those in the broader student population (Bruhn 2020). As such, whether the causal impacts of attending a better school differ by academic advantage across a representative sample of schools and students is unknown. By exploring differences in the effect of attending more effective schools across all schools and all students in a large public school district, we shed light on this issue.

A second motivation for our work is that both economists and social psychologists have found that differences in SED (or noncognitive skills) may explain attainment gaps by gender (Jacob 2002) and socioeconomic status (SES; Claro, Paunesku, and Dweck 2016; Liu 2020). Moreover, experimental studies in psychology find that (*a*) students from low-income families or who are academically lower achieving may benefit from mindset interventions

author, C. Kirabo Jackson, at kirabo-jackson@northwestern.edu. Information concerning access to the data used in this paper is available as supplemental material online.

¹ Angrist et al. (2012), Walters (2018), and Cohodes, Setren, and Walters (2020) find that less advantaged Boston-area charter applicants benefit more from attending oversubscribed charter schools. Conversely, looking at charter-like schools in India, Kumar (2020) finds little difference between more and less advantaged students. Looking at elite school attendance, Dustan, de Janvry, and Sadoulet (2017) and Oosterbeek, Ruijs, and de Wolf (2020) find negative effects in Mexico City and Amsterdam, while Shi (2020) finds the opposite in North Carolina; Barrow, Sartain, and de la Torre (2020) report mixed results in Chicago.

(Sisk et al. 2018) and (b) interventions that promote a sense of belonging are beneficial for the educational outcomes of minoritized (including Black and Latinx) youth (Walton and Cohen 2007; Walton and Cohen 2011; Gray, Hope, and Matthews 2018; Brady et al. 2020; Murphy et al. 2020). If so, measures of school quality that exclude impacts on SED may miss important components of school quality for disadvantaged or minority populations. To shed light on this issue (a novel contribution), we identify school effectiveness explained by test score value-added only versus value-added in other dimensions (i.e., SED and behaviors) and examine these different effects by educational advantage.

We leverage detailed data from Chicago Public Schools (CPS) linking students to high schools and colleges along with test scores, administrative records, and self-reported survey measures of SED over time. First, we categorize students. We use student behaviors, survey measures, and test scores in eighth grade to predict their educational outcomes and create a latent educational advantage index for each student. Students who are low on this index are more likely to be minoritized and male and have low eighth-grade test scores, low socioemotional measures, and more absences and disciplinary incidents than those who are high on the index. Next, we identify schools' impacts on student outcomes (i.e., value-added) by comparing end-of-year outcomes across schools while conditioning on lagged outcomes and other covariates. As in Jackson et al. (2020), we estimate schools' impacts on test scores and socioemotional survey measures in ninth grade. Additionally, we estimate school impacts on behaviors (Heckman and Kautz 2012) and combine effects across all outcomes to create an overall school effectiveness index. Finally, we estimate the effect on educational attainment and school-based arrests of attending a more effective school for students with different levels of educational advantage.

Focusing on heterogeneous effects, all students benefit from attending more effective schools—rejecting a model in which only the most advantaged, or marginal, students benefit from better schools. Looking at the longer-run outcomes, less advantaged students experience larger marginal effects. Specifically, for those in the bottom decile of the distribution, attending a school at the 85th percentile of the effectiveness distribution versus one at the median is associated with a 3.8 percentage point increase in high school graduation, a 3.6 percentage point increase in college going, and a 2.1 percentage point reduction in being arrested. The corresponding estimates for those in the top decile are a 0.5 percentage point increase in high school graduation, a 1.5 percentage point increase in college going, and a 0.22 percentage point reduction in being arrested. These patterns are similar within ethnic groups and by gender, so they are not driven by comparisons across broad demographic groups. We test for whether the pattern of effects reflects disadvantaged students being more likely to be marginal for certain outcomes. This may explain the patterns for arrests and high school graduation but does not explain the patterns for college outcomes.

To help explain these differential impacts by educational advantage, we look at the different components of school effectiveness. While test score value-added and the effectiveness index have similar marginal effects for the most educationally advantaged, for the least advantaged, overall effectiveness predicts much larger effects on longer-run outcomes than test score value-added alone. Additional patterns suggest that this is due, in part, to less academically advantaged students being relatively more responsive to school impacts on SED or soft skills.

Our estimates suggest that the least advantaged students gain the most from attending effective schools. However, in our data the least effective schools are disproportionately attended by the least advantaged students. We simulate a policy that shuts down the least effective schools and reassigns students to other schools. This leads to sizable gains for reassigned students, small negative spillover effects for students at receiving schools, and positive benefits overall—underscoring the likely gains from improving the schools attended by the least advantaged students.

The remainder of the paper proceeds as follows: section II describes the data used, and section III details the methods used and validates our methodology as representing causal impacts. The results are presented in section IV, and section V concludes.

II. Data

As in Jackson et al. (2020), we use administrative data from CPS obtained from the UChicago Consortium on School Research. CPS is a large urban school district. Our data cover six cohorts of ninth graders between the spring of 2012 and 2017, during which there were 150 general education public (neighborhood/charter/vocational/magnet) high schools.² We can estimate effectiveness for 143 of these schools,³ which serve largely ethnically minoritized students (40% Black and 46% Latinx) and economically disadvantaged students (85%).⁴ We include only the first observations for each ninth grader ($n = 160,148$) to remove sample selection biases due to grade repetition. For high school graduation and school-based arrests we focus on three cohorts of ninth graders between 2012 and 2014 ($n = 82,146$), and for college outcomes we focus on two cohorts of ninth graders in 2012 and 2013 ($n = 55,560$).

² CPS offers more than 160 high school programs (<https://www.cps.edu/schools/grade-levels-served/high-school/>). Our count includes only the general education programs.

³ Of the seven excluded schools, five are observed in the data only for a single year—precluding an out-of-sample school estimate—and two do not have enough data to estimate value-added (i.e., fewer than 10 students with survey data and test score). The included schools cover more than 95% of all ninth graders in these cohorts.

⁴ Note that tests were not administered in 2017 and are therefore missing in that year.

because these students are old enough to have attended college.⁵ The data and sample are summarized in table 1 and discussed below.

Survey measures.—Some of our key variables are survey measures of SED. The SED constructs captured by these surveys are hypothesized to be particularly important for the success of disadvantaged youth. Responses are collected by CPS on a survey administered to students in 2008–9 and then every year from 2010–11 onward. These survey items are not part of Chicago’s accountability system, and response rates were high (75%). However, nonresponse was higher for low achievers (table A2). Note that our analysis of impacts on longer-run outcomes is based on all students irrespective of survey completion. Each survey measure was composed of several items, and students responded to each item using point scales to indicate agreement (e.g., 1 = strongly disagree to 4 = strongly agree). Rasch analysis was used to model responses and calculate a score for each student on each construct (for measure properties, see table A3). Following Jackson et al. (2020), we combine the interpersonal-related questions into a social index⁶ and the work-related questions into a work hard index.⁷ To create each index, we take the first predicted principal factor among the survey items. This is a weighted average of the individual items that represents the maximum variance direction in the data (Jolliffe 2002). We then standardize the predicted factors to create indexes that have mean zero and unit variance. All indexes used in this paper adopt this approach and have correlations over 0.95 with the simple mean of standardized items.

Behavior measures.—Motivated by work showing that impacts on behaviors measure skills not well captured by test score impacts (e.g., Heckman, Pinto, and Savelyev 2013; Jackson 2018; Petek and Pope 2020; Liu and Loeb

⁵ See table A1 (tables A1–A9, B1, E1, E2 are available online) for a breakdown of the number of observations for each outcome by cohort.

⁶ Two of the SED survey measures relate to one’s relationship with others in the school. The first is interpersonal skills, and the second is a measure of belonging. Interpersonal skills include the following: I can always find a way to help people end arguments; I listen carefully to what other people say to me; I’m good at working with other students; I’m good at helping other people. Belonging includes the following: I feel like a real part of my school; people here notice when I’m good at something; other students in my school take my opinions seriously; people at this school are friendly to me; I’m included in lots of activities at school.

⁷ Three survey measures capture students’ orientation toward hard work. These are academic effort, the perseverance facet of grit, and academic engagement. Academic effort includes the following: I always study for tests; I set aside time to do my homework and study; I try to do well on my schoolwork even when it isn’t interesting to me; if I need to study, I don’t go out with my friends. Grit includes the following: I finish whatever I begin; I am a hard worker; I continue steadily toward my goals; I don’t give up easily. Academic engagement includes the following: the topics we are studying are interesting and challenging; I usually look forward to this class; I work hard to do my best in this class; sometimes I get so interested in my work I don’t want to stop.

Table 1
Summary Statistics

	Analytic Sample		Bottom Decile of Educational Advantage		Top Decile of Educational Advantage	
	Mean	SD	Mean	SD	Mean	SD
Demographics						
Female	.5013	.5000	.2295	.4205	.7090	.4543
Special education (IEP)	.1816	.3855	.4611	.4985	.0550	.2280
Free lunch	.7748	.4177	.9520	.2138	.4311	.4952
Reduced-price lunch	.0748	.2630	.0168	.1285	.1661	.3721
Census block SES	-.4485	.8732	-.5744	.8141	-.1166	.9166
White	.0896	.2856	.0378	.1908	.2306	.4212
Black	.4043	.4908	.5495	.4976	.2254	.4179
Native American	.0017	.0413	.0024	.0493	.0024	.0493
Asian/Pacific Islander	.0323	.1768	.0014	.0370	.1899	.3922
Latinx	.4584	.4983	.4069	.4913	.3412	.4741
Ninth-Grade Intermediate Outcomes						
Test scores in ninth grade	-.0095	.9913	-1.0053	.6532	1.3844	.7314
Work hard in ninth grade	.0162	1.0014	-.2696	1.0380	.3825	.9819
Social in ninth grade	.0025	.9989	-.2425	1.0456	.3425	.9792
Surveys in ninth grade	.0090	1.0013	-.2993	1.0460	.4080	.9723
Behavior in ninth grade	.1081	.7869	-.5155	1.5421	.3472	.1838
Days absent in ninth grade	14.9988	18.6334	33.6946	27.7080	5.5881	7.7442
Days suspended in ninth grade	.8042	3.2886	2.9187	6.5937	.0653	.6887
Disciplinary incidents in ninth grade	.0769	.4183	.2910	.8709	.0069	.0953
Eighth-Grade Measures						
Math in eighth grade	.1992	.9330	-.8528	.6151	1.7506	.7655
ELA in eighth grade	.2038	.9312	-.8792	.8332	1.5784	.7948
Emotional health in eighth grade	.0703	.8896	-.1918	.8831	.3175	.9211
Academic engagement in eighth grade	.2649	.9067	.1446	.8476	.3423	1.0038
Grit in eighth grade	.0434	.8303	-.3444	.8699	.4596	.7952
School connectedness in eighth grade	.1400	.8937	-.0228	.8594	.4301	.9710
Study habits in eighth grade	.1522	.8833	-.2094	.8455	.6695	.9512
Absences in eighth grade	8.7060	8.5615	19.2280	12.4523	4.6015	3.8142
GPA in eighth grade	2.8017	.7776	2.0370	.7739	3.6012	.4825
Days suspended in eighth grade	.4366	1.8074	2.2078	4.3782	.0177	.2514
Incidents in eighth grade	.0639	.3329	.3702	.8326	.0005	.032
Long-Term Outcomes						
Any school-based arrest	.0372	.1891	.1236	.3291	.0047	.0684
Graduation	.7406	.4383	.4333	.4956	.9334	.2493
Enrolled in any college within 2 years	.5322	.4990	.1842	.3877	.8615	.3455
Enrolled in a 4-year college within 2 years	.3432	.4748	.0722	.2588	.7638	.4248
Enrolled in a 2-year college within 2 years	.2761	.4471	.1280	.3342	.2514	.4338
N	160,148		16,015		15,995	

NOTE.—Number of observations may vary by variable because of missingness and variation in cohorts for which a variable was collected. For more information, see table A1. Note that because the sample size is not perfectly divisible by 10 and because observations with the same value of educational advantage are placed in the same decile group, the decile groups are slightly unequal in size. GPA = grade point average; IEP = individualized education program.

2021), we augment the survey-based measures used in Jackson et al. (2020) and also include student behaviors from CPS administrative data. These include the number of excused and unexcused absences, severe disciplinary incidents (eligible for suspension), and days a student is suspended, in each grade. In the analytic sample with nonmissing behaviors ($n = 157,628$), the average ninth grader is absent 15 days and suspended 0.8 days. Approximately 7.7% of these are involved in a severe disciplinary incident. We summarize these three ninth-grade measures by forming a behaviors index, which is the standardized first predicted principal factor.

Test score measures.—The “hard” skills measure in our data is standardized test scores. To allow for comparability across grades, test scores were standardized to be mean zero unit variance within grade and year among all CPS test takers. Tests were not administered to ninth graders in spring 2017, so there is nontrivial missingness for this ninth-grade outcome. However, table A2 shows that the sample of test takers is similar to the full sample. We form a test score index using the standardized first predicted principal factor among standardized math and English scores.

Long-run outcomes.—A key longer-run outcome is having a school-based arrest (among those old enough to have graduated high school). These are arrests for activities conducted on school grounds, those occurring during off-campus school activities, or those due to a referral by a school official. During our sample period, 3.7% of first-time ninth graders, 5.2% of males, and 7.7% of Black males had a school-based arrest. Roughly 20% of juvenile arrests in 2010 were school-based arrests (Kaba and Edwards 2012), so these have important long-term implications. Our other longer-term outcomes include high school graduation and enrollment and persistence in college. High school completion is obtained from school leaving files from the years 2010 through 2018. We define a student as having graduated high school if they are marked as leaving high school because they graduated. Note that this definition does not include equivalency credentials, such as a GED. About 74% of first-time ninth graders in CPS graduate high school. Our second key long-run outcome is enrollment in college. Our college data come from the National Student Clearinghouse (NSC) and are merged with all CPS graduates. We code a student as enrolling in college if they are observed in the NSC data within 2 years of expected high school graduation (2012 and 2013 cohorts only). Students who did not graduate high school are coded as not enrolling in college, so our outcomes indicate “graduating high school and enrolling in college.” Because graduating high school is a prerequisite for college enrollment, these two definitions are largely the same.⁸ About 53% of first-time ninth graders enrolled in college. We further divide college enrollment into 2-year and 4-year college. In

⁸ A limitation of our data is that it will miss students who enroll as ninth graders but who graduate from a high school outside of CPS. This is relatively rare, particularly for the populations for which we find the most pronounced effects.

our sample, 34% and 28% of students enroll in a 4-year or 2-year college within 2 years of expected graduation, respectively.

III. Methods

Our analysis involves three steps. First, we use eighth-grade data to classify students as more or less advantaged. Second, we estimate school value-added on several ninth-grade measures and combine them to create an overall school effectiveness measure. Finally, we estimate the effect of attending a more effective school among students of differing educational advantage.

A. Classifying Students

To classify students along a single dimension, we rank students by their likelihood to attain more years of education. We refer to students who are more likely to attain more years of education (based on observed characteristics before entering high school) as more educationally advantaged. Our approach exploits the fact that we have a rich set of observable characteristics that may predict educational attainment and that we observe multiple measures of educational attainment.

The basic idea is that some underlying educational advantage, y_i^* , is a linear function of observable characteristics X so that $y_i^* = X\pi + \varepsilon$. Individuals with higher levels of educational advantage attain higher levels of education, where there are some unobserved thresholds between education levels. That is, for all individuals i

$$y_i = \begin{cases} \text{no high school} & y_i^* \leq \chi_1, \\ \text{graduate high school} & \chi_1 > y_i^* \leq \chi_2, \\ \text{attend 2-year college} & \chi_2 > y_i^* \leq \chi_3, \\ \text{attend 4-year college} & y_i^* > \chi_3. \end{cases}$$

The probability of observing outcome $y_i = k$ is then $\Pr(y_i = k) = \Pr(\chi_{k-1} < X\pi \leq \chi_k)$. The probability of observing the data is the product of these probabilities across all individuals i . Assuming a normally distributed error term, we solve for the set of estimates $(\hat{\pi}, \hat{\chi}_{k=1}, \hat{\chi}_{k=2}, \hat{\chi}_{k=3})$ that are most consistent with the observed data by estimating an ordered probit model by maximum likelihood.

Our predictors of the education outcomes include measures of lagged test scores (quadratics of eighth-grade math and English language arts [ELA]), eighth-grade survey measures, and lagged behaviors.⁹ We also include demographics (lunch status, race, gender, and interactions between race and

⁹ Because these variables do not have a lot of variation in early grades, we include an indicator for being in the top quartile of absences in eighth grade and an indicator for having any severe disciplinary incidents in seventh or eighth grade.

gender). Once the parameter estimates have been estimated, we take the fitted values of the latent variable, $\hat{LA}_i = X\hat{\pi}$, as our estimated latent educational advantage. Note that we use leave-year-out models to avoid mechanical correlation between our predicted and actual education levels for each student i . However, to show the relationship between the advantage index and the observable covariates, we present the coefficient estimates from the ordered probit model for all students in the 2012 and 2013 cohorts in table A4.

1. Differences in Incoming Attributes and Outcomes by Educational Advantage

We present summary statistics for the top and bottom deciles of the educational advantage distribution in the middle and right panels of table 1. The top decile contains almost three times more females than the bottom decile (71% vs. 23%), about eight times fewer students in special education (5.5% vs. 46%), and less than half the share of students who qualify for free lunches (43% vs. 95%). The top decile has more white students than the bottom decile (23% vs. 3.8%) as well as more Asian students (19% vs. 0.14%), but fewer Latinx students (34.1% vs. 40.7%) and Black students (22.5% vs. 55%). Relative to those in the bottom decile, students in the top decile have eighth- and ninth-grade test scores more than 2 standard deviations higher, have fewer absences (5.6 vs. 33.7 days) and days suspended (0.065 vs. 2.9 days), and are involved in fewer severe incidents (0.007 vs. 0.29) in ninth grade.

To illustrate the differences in our main longer-run outcomes by advantage, we plot the average of our key outcomes for each percentile of the index in figure 1. At the bottom of the index (the bottom 20%), even though about 40% of students graduate from high school, few (about 17%) go to any college and even fewer (10%) attend a 4-year college. In the middle of the distribution (between the 40th and 60th percentiles), the high school graduation rate is about 75%, the college-going rate is about 50%, and both the 4-year and the 2-year college-going rates are around 25%. As one looks to the top of the distribution (the top 20%), the high school graduation rate is above 90%. Interestingly, the 4-year college-going rate increases to about 70%, while the 2-year college-going rate remains at 25%. That is, as one goes up the educational advantage distribution, 4-year college going increases but 2-year college going does not. School-based arrests are concentrated among students with very low educational advantage. In the bottom decile the arrest rate is a sizable 12.4% (see table 1), while for those above the median it is almost zero (0.02%). Even though educational advantage varies by ethnicity and gender, figure A1 (figs. A1–A3, C1, C2, D1, D2, E1, E2 are available online) shows similar relationships between outcomes and educational advantage within groups. As such, the heterogeneity analysis is not merely based on comparisons across demographic groups.

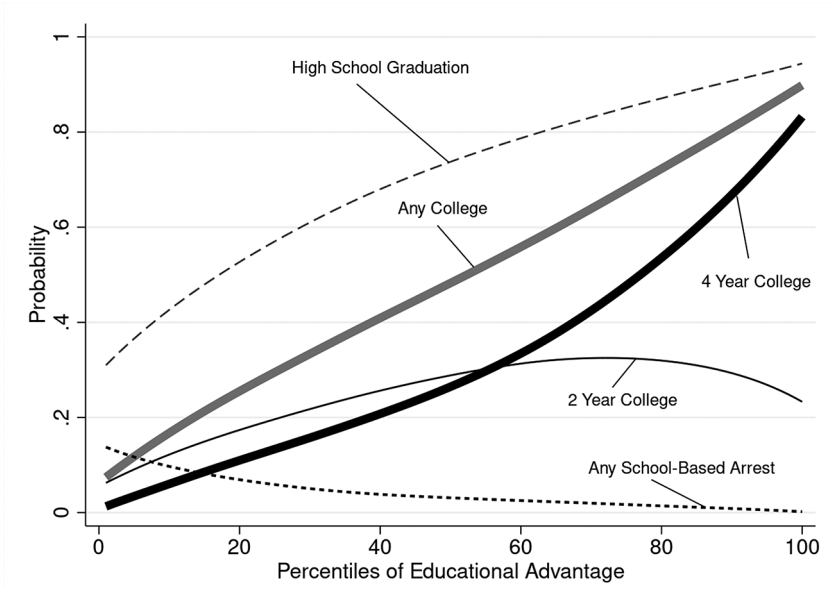


FIG. 1.—Average outcomes, by estimated educational advantage. This figure plots the average of each outcome for different percentiles of the estimated educational advantage distribution. The predicted educational advantage is the fitted value from an ordered probit model predicting the level of education attained based on all eighth-grade measures and demographics (in all other years). We present the coefficient estimates from the ordered probit model for the full sample in table A4. We also present plots of outcome educational advantage within race and gender groups in figure A2.

B. Classifying Schools

We use value-added models to estimate schools’ impacts on ninth-grade SED, behaviors, and test scores. We then combine the value-added estimates across outcomes to form an overall school effectiveness index. We present evidence that these impacts can be interpreted causally.

1. Identifying School Impacts on SED, Behaviors, and Test Scores

We seek to isolate the causal effects of individual schools (i.e., value-added) on student measure $q \in Q = \{\text{test scores, work hard, social, behaviors}\}$. School j ’s value-added on measure q reflects how much school j increases measure q between eighth and ninth grade relative to the changes observed for similar students (based on lagged outcomes and demographics) who attended different schools. We model the ninth-grade measure q of student i who attends school j with observable characteristics Z_{ijt} in year t as

$$q_{ijt} = \underbrace{\beta_q Z_{ijt}}_{\text{effect of observables}} + \underbrace{\tau_{t,q}}_{\text{cohort fixed effect}} + \underbrace{\alpha_{i,q} + \varepsilon_{ijt,q}}_{\text{combined error } v_{ijt,q}}, \tag{1}$$

where Z_{ijt} includes lagged measures (i.e., eighth-grade test scores, surveys, behaviors), gender, ethnicity, free-lunch status, and the SES of the student's census block and $\tau_{t,q}$ is a cohort fixed effect.¹⁰ To account for correlation between schools and covariates, we follow Chetty, Friedman, and Rockoff (2014) and include a school-specific intercept, $\alpha_{j,q}$, so that β_q is estimated using variation across students at the same school, and $\varepsilon_{ijt,q}$ is a within-school student-level error.¹¹ Because β_q is estimated using time-varying within-school variation in individual student attributes, $\alpha_{j,q}$ may reflect the effect of persistent differences in peer demographics across schools. Where $v_{ijt,q} = \alpha_{j,q} + \varepsilon_{ijt,q}$ is the true combined error, $u_{ijt,q}$ is the empirical combined residual. The school-year average combined residuals from this regression is our estimated impact on measure q of attending a school in a given year, $\hat{\theta}_{jt,q}^{\text{VA}}$. Formally, where N_{jt} is the number of students attending school j in year t ,

$$\hat{\theta}_{jt,q}^{\text{VA}} = \frac{\sum_{i \in jt} u_{ijt,q}}{N_{jt}}. \quad (2)$$

When using value-added to predict outcomes for a particular cohort, we exclude data for that cohort when estimating value-added to avoid mechanical correlation. To aid precision, we follow Chetty, Friedman, and Rockoff (2014) and use value-added with drift, which places more weight on value-added for other years that are more highly correlated with the prediction year.¹² Our leave-year-out predictor for measure q in year t is $\hat{\mu}_{jt,q}$ from equation (3), where the vector of weights $\hat{\psi}_q = (\hat{\psi}_{t-l,q}, \dots, \hat{\psi}_{t-1,q}, \hat{\psi}_{t+1,q}, \dots, \hat{\psi}_{t+l,q})'$ are selected to minimize mean squared forecast errors:

$$\hat{\mu}_{jt,q} = \sum_{m=t-l}^{t-1} \hat{\psi}_{m,q} [\hat{\theta}_{jm,q}^{\text{VA}}]. \quad (3)$$

The term $\hat{\mu}_{jt,q}$ is our best prediction based on other years of how much that school will increase measure q between eighth and ninth grade relative to the improvements of similar students at other schools. We use leave-year-out predictions for all analyses but for brevity refer to them simply as value-added.

2. Correlations across Effects on Different Measures

To assess whether each value-added measure represents impacts on a different or the same dimension of school quality, we correlate school impacts across these four measures. For these cross-sectional correlations only, we

¹⁰ The census block SES measure is the average of occupation status and education levels in the block.

¹¹ Results are similar using models that exclude $\alpha_{j,q}$ and include school-level averages of the individual covariates.

¹² If all years' value-added were equally predictive of outcomes in year t , then the best leave-year-out predictor for a school would be the average value-added for that school in all other years. However, adjacent years tend to be more highly correlated with one another than less temporally proximate years (see the top panel of table 2).

need only one observation per school. Accordingly, our value-added measure for each school in all years, $\hat{\theta}_{j,q}^{VA}$, uses the full sample and is given by equation (4), where N_j is the number of students assigned to school j across all years:

$$\hat{\theta}_{j,q}^{VA} = \frac{\sum_{it \in j} (u_{ijt,q})}{N_j} \equiv \underbrace{\theta_{j,q}^{VA}}_{\text{real value-added on outcome } q} + \underbrace{\xi_{j,q}}_{\text{estimation error}}.$$

Where q_1 and q_2 connote different outcomes, we report the raw pairwise correlations between the value-added for different outcomes (i.e., $\text{Corr}(\hat{\theta}_{j,q_1}^{VA}, \hat{\theta}_{j,q_2}^{VA})$ in the middle panel of table 2). Because each of these value-added is measured with error, the true correlations could be higher than this (due to attenuation bias from random estimation errors) or lower than this (if the measurement errors are correlated across outcomes in the same year). Accordingly, we follow Beuermann et al. (2023) and implement a split-sample approach that uncovers the real correlation between effects across outcomes under

Table 2

Temporal Stability of Value-Added and Correlations across Value-Added

	Test Score Value-Added	Social Value-Added	Work Hard Value-Added	Behavior Value-Added
Correlations of Value-Added within Outcomes across Time				
$t + 1$	.4522	.5586	.3847	.6022
$t + 2$	.1884	.3780	.2158	.5338
$t + 3$	.4012	.3425	.2303	.4801
$t + 4$	.4674	.3308	.2937	.3475
Correlations of Average School-Level Value-Added across Outcomes (143 Schools)				
Test scores value-added	1			
Social value-added	.551	1		
Hard work value-added	.190	.397	1	
Behavior value-added	.362	.184	.055	1
Disattenuated Correlations of Average School-Level Value-Added across Outcomes (143 Schools)				
Test scores value-added	1.000			
Social value-added	.852	1.000		
Hard work value-added	.322	.780	1.000	
Behavior value-added	.438	.259	.084	1.000

NOTE.—All reported results are restricted to school-year cells with at least 10 respondents. The top panel reports, for each ninth-grade measure, the correlations between a school's value-added in year t and value-added for years $t + 1$, $t + 2$, $t + 3$, and $t + 4$. The bottom panel reports the correlations between the value-added (estimated across all years) for the ninth-grade measures.

the assumption that estimation errors are unrelated over time both within and across outcomes, as in Kane and Staiger (2008) and Jackson (2013).¹³

We report the disattenuated correlations in the bottom panel of table 2. The patterns are similar to those of the raw correlations. Test score value-added is strongly related to the social dimension of SED value-added ($\rho = 0.85$) and has moderate correlations with work hard value-added ($\rho = 0.32$) and behaviors value-added ($\rho = 0.44$)—suggesting some underlying dimension of school quality that is associated with higher value-added in all of these dimensions. Another interesting pattern is that behaviors value-added is more strongly related to test score value-added than the survey-based SED value-added—suggesting that surveys and behaviors may measure different dimensions of socioemotional skills. While the low correlation between the survey- and behavior-based measures of socioemotional skills is interesting, we leave exploration into why to future work.

3. Creating an Overall School Effectiveness Index

To better understand correlations in the data, we conduct factor analysis of the school effects (table A5). The model finds that a single underlying factor is positively related to all of the value-added and explains almost all of the common variation in these value-added. As such, we combine our value-added (work hard, social, test scores, and behaviors) into a single index of school effectiveness. Our overall index is the predicted first principal factor of these four variables. The overall index, $\hat{\omega}_{jt}$, is a weighted average of the different value-added estimates given by equation (5):¹⁴

$$\hat{\omega}_{jt} = (0.3)\hat{\mu}_{jt,\text{test scores}} + (0.34)\hat{\mu}_{jt,\text{work hard}} + (0.26)\hat{\mu}_{jt,\text{social}} + (0.1)\hat{\mu}_{jt,\text{behaviors}} \quad (5)$$

¹³ If measurement errors are uncorrelated over time, then the correlation between the value-added estimated using data only during the even years for one outcome ($\hat{\theta}_{j,\text{even},q_1}^{\text{VA}} = \theta_{j,\text{even},q_1}^{\text{VA}} + \zeta_{j,\text{even},q_1}$) and those using data only during the odd years for the other outcome ($\hat{\theta}_{j,\text{odd},q_2}^{\text{VA}} = \theta_{j,\text{odd},q_2}^{\text{VA}} + \zeta_{j,\text{odd},q_2}$) would not be biased by correlated errors across outcomes because $\text{Corr}(\zeta_{j,\text{even},q_1}, \zeta_{j,\text{odd},q_2}) = 0$. However, this even-odd cross outcome correlation ($\hat{\rho}_{12}^{\text{even-odd}}$) will be attenuated by random estimation errors. When two variables are measured with random errors, the raw correlations between the two variables (in this case, $\hat{\rho}_{12}^{\text{even-odd}}$) reflect the true correlation times the square root of the product of the reliability of each outcome (Spearman 1904). In this case, that is $\hat{\rho}_{12}^{\text{even-odd}} = \rho_{12}^{\text{even-odd}}(R_{\text{even},q_1}R_{\text{odd},q_2})^{1/2}$, where R_{even,q_1} is the reliability of $\hat{\theta}_{j,\text{even},q_1}^{\text{VA}}$ and R_{odd,q_2} is the reliability of $\hat{\theta}_{j,\text{odd},q_2}^{\text{VA}}$. As such, one can disattenuate the raw correlation by dividing by the square root of the product of the reliability ratios for each measure (Spearman 1904). The reliability of each measure can be obtained using the correlation between even and odd year estimates for the same outcome (i.e., $\hat{\rho}_{11}^{\text{even-odd}}$ and $\hat{\rho}_{22}^{\text{even-odd}}$). With the clean raw correlations ($\hat{\rho}_{12}^{\text{even-odd}}$) and the estimated reliability ratios ($\hat{\rho}_{11}^{\text{even-odd}}$ and $\hat{\rho}_{22}^{\text{even-odd}}$), we compute clean disattenuated correlation estimates $r_{12} = [\hat{\rho}_{12}^{\text{even-odd}}]/(\hat{\rho}_{11}^{\text{even-odd}}\hat{\rho}_{22}^{\text{even-odd}})^{1/2}$.

¹⁴ The weights have been normalized to sum to 1 for ease of interpretation.

We standardize the quality index to be mean zero, unit variance. As we show in section IV.C, the index is generally a better predictor of school impacts on longer-run outcomes than the value-added on the individual measures.

4. *Some Measurable Differences between More and Less Effective Schools*

To describe more or less effective schools, we present binned scatterplots of observable school attributes by 20 ventiles of school effectiveness in figure 2. Despite a strong correlation between average outcomes and student demographics, among the least and most effective schools there are small differences in the proportion of nonwhite students (0.95% vs. 0.93% in the top and bottom ventiles, respectively) and those receiving free and reduced-priced lunch (0.93% vs. 0.91% in the top and bottom ventiles, respectively). This echoes results from randomized lotteries in New York and Denver (Angrist et al. 2022). The more effective schools tend to be larger and have slightly larger class sizes. Noble charter schools are overrepresented among the most effective schools (consistent with lottery-based studies; e.g., Davis and Heller 2019), and selective-enrollment schools are somewhat more likely to be among the most effective schools (consistent with Barrow, Sartain, and

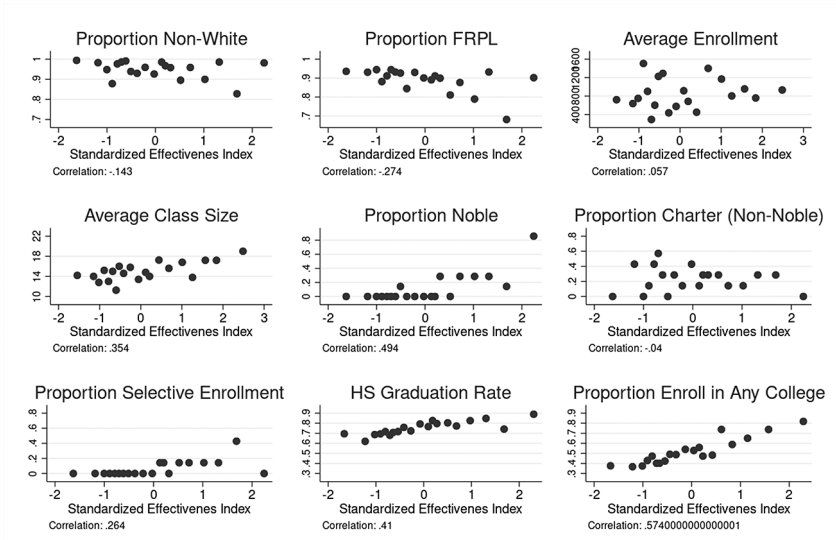


FIG. 2.—Difference in characteristics among most and least effective schools. Each panel is a binned scatterplot (20 bins) of the school characteristic against the standardized estimated school effectiveness. We report the raw correlation between the school characteristic and effectiveness below each plot for a sample of 131 schools. Share FRPL is defined as the share of students eligible for free or reduced-price lunch at a school. Note that class size and enrollment data come from the Illinois report card for the 2016–17 school year. Data are available at <https://www.illinoisreportcard.com>.

de la Torre [2020], who find positive effects on college going using a regression discontinuity design). More effective schools also have better high school graduation and college-going rates on average.

C. Estimating School Effectiveness Impacts by Educational Advantage

To quantify the effect of attending a school with 1 standard deviation higher predicted overall effectiveness, we regress each outcome on the standardized school effectiveness index (plus controls). Specifically, where Y_{ijt} is an outcome and $\hat{\omega}_{jt}$ is the standardized out-of-sample predicted effectiveness, we estimate the following model by ordinary least squares (OLS):

$$Y_{ijt} = \delta \hat{\omega}_{jt} + \beta_1 Z_{ijt} + \tau_t + \varepsilon_{ijt}. \quad (6)$$

All variables are as defined above, and τ_t is a year fixed effect. In addition to the controls in equation (1), we include school-level averages of all individual attributes. Standard errors are clustered at the school level.¹⁵ To flexibly present differences in the marginal impacts by student type, we estimate equation (6) separately for each decile d of the estimated educational advantage index, and we plot the decile-specific marginal effects δ_d against the educational decile d (along with confidence intervals for each decile-specific effect). We formally test the hypothesis that the marginal effects are linearly related to educational advantage by pooling the data for all students and adding an interaction between the estimated advantage decile and school effectiveness ($\widehat{LA}_i \times \hat{\omega}_{jt}$) and fixed effects for each educational advantage decile (α_d), as in equation (7):

$$Y_{ijt} = \delta_1 \hat{\omega}_{jt} + \delta_2 (\widehat{LA}_i \times \hat{\omega}_{jt}) + \beta_1 Z_{ijt} + \tau_t + \alpha_d + \varepsilon_{ijt}. \quad (7)$$

Under no linear relationship between the marginal effect of schools and educational advantage, the slope between effectiveness and the marginal effect (i.e., δ_2) is zero.¹⁶

D. Testing for Selection

We first present our preferred average model. Columns 1, 5, and 9 in table 3 report the coefficient on the standardized effectiveness index in a

¹⁵ Individuals with missing eighth-grade surveys or test scores are given imputed values. We regress each survey measure or test score on all observed pre-eighth-grade covariates. We then obtain predicted eighth-grade values on the basis of these regressions and replace missing values with the predictions. Results are similar with and without imputation.

¹⁶ To assuage concerns that the advantage is a generated regressor, as a robustness check we estimate eq. (7) but instrument for the interaction ($\widehat{LA}_i \times \hat{\omega}_{jt}$) with eighth-grade covariates (a test score index, a surveys index, and a behaviors index), each interacted with estimated effectiveness. The results are very similar (see table A11).

Table 3
Testing for Selection

	Intermediate Outcomes				Main Longer-Run Outcomes				College Outcomes			
	Ninth-Grade Test Scores		Predicted		High School Graduation		Predicted		Enrolled in 4-Year College within 2 Years		Predicted	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
School effectiveness index	.0890*** (.0133)	.0560*** (.0177)	.0527*** (.0137)	-.000275 (.00430)	.0241*** (.00430)	.0362*** (.0121)	.0171** (.00845)	.00266 (.00183)	.0416*** (.00903)	.0355*** (.0143)	-	.00347 (.00467)
Observations	102,235	102,235	16,386	102,235	82,146	82,146	8,188	82,146	55,564	55,564	3,399	55,564
F-statistic on first stage		321				393				374.9		
Number of families			7,786				3,902				1,634	
In-School Arrests												
School effectiveness index	.102*** (.00755)	.105*** (.0186)	.0586*** (.0173)	.00260 (.00852)	-.00804*** (.00271)	-.0132** (.00622)	-.00957* (.00501)	-.000713 (.000993)	-.00570 (.00456)	-.0211* (.0120)		-.000159 (.00213)
Observations	120,129	120,129	27,103	120,129	82,146	82,146	8,188	82,146	55,564	55,564	3,399	55,564
F-statistic on first stage		294.7				393				374.9		
Number of families			13,584				3,902				1,634	

	Ninth-Grade Behaviors			Enrolled in Any College within 2 Years		Persist in College after 1 Year			
School effectiveness index	.0571*** (.0127)	.0702*** (.0207)	.0317*** (.0112)	.00552 (.00867)	.0257*** (.00654)	.00259 (.00247)	.0203*** (.00550)	.0153 (.0139)	.00235 (.00230)
Observations	157,628	157,628	41,711	157,628	55,564	55,564	55,564	55,564	55,564
F-statistic on first stage		319.5			374.9			374.9	
Number of families			19,420			1,634			1,634
Sibling fixed effects			X			X			X
School assignment IV		X			X			X	

NOTE.—Results are based on regression of outcomes on out-of-sample school effectiveness. All models include individual demographic controls (race/ethnicity, free and reduced-price lunch, and gender), eighth-grade lags (math and ELA test scores, survey measures, absences, and discipline), and school-level averages for all of the demographics and lagged measures, as well as year fixed effects. We also include the SES of the student census block proxied by average occupation status and education levels. Missing eighth-grade measures were imputed using seventh-grade measures and demographic characteristics. For the longer-run college outcomes, the sample includes two cohorts of first-time ninth graders in spring 2013 and 2014. For the longer-run high school outcomes, the sample includes three cohorts of first-time ninth graders in spring 2012, 2013, and 2014. For the measures, the sample includes six cohorts of first-time ninth graders between spring 2012 and 2017. Columns 4, 8, and 12: predicted outcomes are fitted values from a linear regression of said outcome on all observed controls. The predictors include lagged measures (i.e., eighth-grade test scores, surveys, and behaviors), gender, ethnicity, free-lunch status, and the SES of the student's census block. To avoid mechanical correlation, we use leave-year-out predicted outcomes (i.e., predicted outcomes based on the relationship between the outcome and covariates in other years). Whether the predictions use relationships in sample or out of sample, the results are the same. The reported point estimates are those on predicted outcomes on the value-added while controlling only for school-average math scores, school-average absences, and the percent white at the school. Note that sample sizes may differ across outcomes because of some missingness in ninth-grade test scores and surveys. Robust standard errors adjusted for clustering at the school level are in parentheses.

* $p < .10$.
 *** $p < .05$.
 *** $p < .01$.

regression of various outcomes on the index and controls for the full sample (from eq. [6]). To take these estimates as reflecting causal impacts requires that, on average, there are no unobserved differences in the determinants of outcomes between students who attend more and less effective schools. However, because students are not randomly assigned to schools, unaccounted-for selection could (a) lead to a spurious correlation between the effectiveness index and outcomes and (b) yield patterns that suggest heterogeneity even when there is none. While there is no way to prove that the effectiveness estimates are unrelated to unobserved determinants of outcomes, we present several tests to show that this is likely satisfied in our setting. We discuss these tests in detail in appendix B (apps. A–E are available online) and provide a brief overview here.

Altonji and Mansfield (2018) and Agrawal, Altonji, and Mansfield (2019) show that when individuals sort into groups, group-level averages can serve as a control function to remove the across-group variation in both observable and unobservable individual characteristics. Accordingly, our main model includes school-level averages of all individual characteristics. To show that this approach is plausible, we show that conditional on only three school-level average variables (school average math scores, absences, and share white) and no other controls, predicted outcomes based on all individual-level characteristics are unrelated to school effectiveness (cols. 4, 8, and 12 in table 3).¹⁷

In another test, to account for the possible selection of students (within families) to schools outside their residentially zoned school, we instrument for the effectiveness of the attended school with the effectiveness of the residentially zoned school (cols. 2, 6, and 10 in table 3).¹⁸ If our results were driven by such selection, our instrumental variable estimates would be zero. For the short-run outcomes (test scores, surveys, and behaviors), the two-stage least square (2SLS) estimates are positive, of the same order of magnitude as the OLS estimates (in cols. 1, 5, and 9), and all statistically significant at the 1% level. For the main long-run outcomes (arrests, high school graduation, college going, and 4-year college going), the point estimates are positive and of the same order of magnitude as the OLS estimates, and three out of four are statistically significant at the 5% level. The results suggest no selection to schools outside of students' residentially zoned schools.

As a final test for selection on unobservables, we compare students from the same family by including sibling fixed effects in the regression models. This removes any bias due to certain kinds of families attending better schools. Because there were very few sibling pairs with both siblings being old enough

¹⁷ While our test is similar to that in Altonji, Elder, and Taber (2005) and Oster (2019), our control function method is justified by theory and includes a very small number of variables relative to the full set of individual predictors of student outcomes.

¹⁸ In Chicago, almost two-thirds of children attend schools other than their zoned school (Hing and Jenniver 2019).

to have college outcomes, we can estimate only the within-family model for the high school outcomes. For these outcomes (cols. 3 and 7), the within-family models are similar to the OLS and 2SLS models, and most (4 out of 5) are statistically significant at the 5% level—suggesting no selection.

If our estimates were biased by selection, strong predictors of outcomes would likely be related to our estimated value-added (conditional on a small number of controls)—but this is not the case. If our results were driven by selection to school assignment zones across families, it would bias our 2SLS results but not our sibling results. If our results were driven by selection to schools within families (and assignment zones), it would bias our sibling results but not our 2SLS results. The similarity between the OLS results, the 2SLS results, and the within-family results (coupled with the lack of selection on observables) suggests that neither is biased. In light of validation work showing that value-added estimates tend to be largely unbiased (e.g., Deming et al. 2014; Angrist et al. 2020), together they show that our estimated school impacts likely reflect causal impacts.

IV. Results

A. Impacts of School Effectiveness on Short-Run Measures: By Advantage

We now consider how school effectiveness effects vary by students' ex ante educational advantage. We estimate these same regressions for each measure and for each decile. To summarize these 10 regressions per outcome, we plot the point estimate and 95% confidence interval for each estimate in figure 3. Table A7 reports the point estimates for individual deciles, table A8 reports the point estimates for the top and bottom three deciles, and table A9 reports the point estimate on the interaction term of interest from equation (7). We also plot the linear relationship between the estimated effect and the educational advantage decile along with the p -value associated with the null hypothesis of no linear relationship between educational advantage and the benefit of attending a more effective school.

The top left panel of figure 3 shows the effect of attending a school 1 standard deviation higher in school effectiveness on students' ninth-grade test scores. All students benefit from attending a more effective school, with larger effects for the less advantaged. The p -value associated with the hypothesis that the effect varies linearly with educational advantage is .07—suggestive of larger marginal benefits for the least educationally advantaged. We now turn to the survey measures. The middle panel shows the effect of attending a school 1 standard deviation higher on the effectiveness index on students' socioemotional measures in ninth grade by decile of advantage. As with test scores, all students benefit from attending a more effective school. However, for surveys, one cannot reject that the impacts are the same throughout the educational advantage distribution. We report the results for ninth-grade

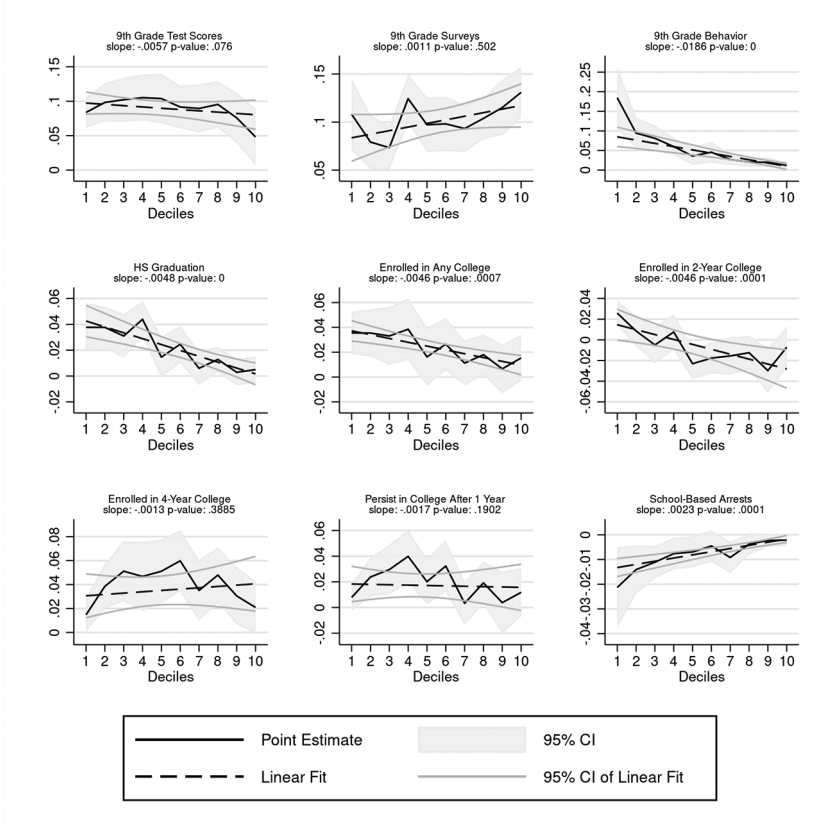


FIG. 3.—Impacts on outcomes, by estimated educational advantage. Each graph represents the marginal impact of a 1 standard deviation increase in overall school effectiveness for different deciles of the educational advantage distribution for a single outcome. Each panel presents the results of 10 separate regressions, each defined as in equation (6). The 95% confidence interval (CI) for each point estimate is depicted by the gray area. The dashed black line in each panel depicts the line of best fit for the relationship between deciles of educational advantage and the marginal effect, including a 95% CI. The reported slope is the coefficient on the interaction between educational advantage and overall effectiveness, δ_2 , from equation (7). We also report the p -value associated with the null hypothesis that this slope is zero (i.e., that there is no heterogeneity).

behaviors in the right top panel of figure 3. Effective schools have the strongest effect on the observed behaviors of less advantaged students. Unlike the survey and test score measures, one rejects the hypothesis of the same marginal effect for all advantage groups at the 1% level. One interpretation of this pattern is that schools have heterogeneous effects on students across the distribution. However, the small impacts for students at the top of the distribution

may be driven by a lack of variation among these students. Specifically, students in the top decile are very unlikely to be involved in a disciplinary incident (0.007 compared with 0.29 in the bottom decile) and have a low absence rate (5.6 days compared with 34 days in the bottom decile), so there is relatively little room for improvement. We show that this may explain the seemingly heterogeneous impacts on the behaviors index in section IV.A.1.

1. Testing for Mechanical Heterogeneity

The basic idea of this test is that for most models of binary outcomes (such as a logit or probit), with the same change in underlying skills the marginal effects will be largest for groups that are marginal (with probability of success close to 0.5) and smallest for groups with probabilities of success farthest from 0.5 (i.e., 0 or 1).¹⁹ Under purely mechanical heterogeneity (i.e., the same underlying change in latent disposition across groups), there will be (a) a negative relationship between the absolute value of the marginal effect and the absolute difference between the group success rate and 0.5 and (b) a predicted marginal effect of zero for groups with an absolute difference from 0.5 of 0.5. We assume a linear relationship, for simplicity, and test for this using the regression in equation (8):

$$|\delta_g| = \alpha + \pi \times (|p_g - 0.5|) + v_g. \quad (8)$$

Where $|\delta_g|$ is the absolute value of the marginal effect for decile group g and p_g is the average success rate for decile group g , then π represents the relationship between the marginal effect and the distance between the baseline success rate for a group and 0.5.

Conducting this test for some of the binary outcomes underlying the behaviors provides strong evidence of such “mechanical” effects (see fig. C1). For the likelihood of being chronically absent, having any suspensions, or having any disciplinary infractions, one rejects that the slope is zero at the 1% level, and one cannot reject that the effect is zero for those who are very likely or unlikely to have a success. To explore whether the heterogeneity is all mechanical or reflects some real heterogeneity, we simulate the distribution of linear slopes (as in fig. 3) under mechanical heterogeneity alone (see fig. C2).²⁰ The actual slopes observed are within the simulated range expected under pure mechanical heterogeneity for most of the binary outcomes comprising the behaviors index. These additional tests are consistent with the similar effect on the continuous measures (i.e., test scores and surveys) across

¹⁹ We expand on the logic of this test formally in app. C.

²⁰ That is, we implement a probit model using the actual data, estimate each individual’s latent disposition, add a constant marginal treatment effect to each latent disposition, and then feed this back into the probit function to simulate “fake” binary treatments driven by mechanical heterogeneity. We do this 1,000 times to simulate the distribution of slopes one would observe due to mechanical heterogeneity alone.

the educational advantage distribution—suggesting little differential effect on underlying skills in the short run.

B. Impacts on Longer-Run Outcomes: By Advantage

We now examine similar figures for the longer-run outcomes (the middle and lower panels of fig. 3). Looking at high school graduation, the marginal impacts of school effectiveness are much larger for less advantaged students. One rejects that the linear relationship between the marginal effect and educational advantage is zero at the 1% level. Indeed, for those in the bottom decile, a 1 standard deviation increase in effectiveness increases high school completion by 3.8 percentage points ($p < .01$), compared with only 0.5 percentage points ($p > .10$) in the top decile. Relative to each groups' baseline level, this is about a 9% increase for those at the bottom of the distribution compared with a 0.6% increase for the top. The differences in the changes in graduation rates across groups are real and economically meaningful. However, one may wonder whether this pattern is due to more students at the bottom being on the margin of high school graduation. We assess this using the test detailed in section IV.A.1. The results of this test are summarized in figure C1, which plots the linear relationships between the absolute value of the marginal effects against the absolute deviation of the group's baseline success rate from 0.5. For high school graduation, the p -value on the slope is significant at the 1% level, and the implied effect is near zero for groups that are very likely or unlikely to have a success. Also, in figure C2, the actual slope is within the distribution of simulated slopes under mechanical heterogeneity—suggesting that the observed heterogeneity in high school graduation is largely mechanical.

Next, we examine enrolling in any college (2 or 4 year) within 2 years of expected high school completion. There are benefits for all groups. However, there are larger increases at the bottom of the distribution, and these differences are statistically significant (the p -value on the slope relating the marginal effect and educational advantage is .0007). More specifically, for the bottom decile, a 1 standard deviation increase in effectiveness increases college going by 3.55 percentage points ($p < .01$), compared with about 1.5 percentage points ($p > .1$) in the top decile. Given the large differences in base rates, the differences in relative marginal impacts are sizable. For the bottom decile, a 1 standard deviation increase in effectiveness increases college going by about 19%, compared with under 2% in the top decile. Unlike for high school graduation, the heterogeneous effect on college going is unlikely to be driven by differences in baseline success rate. The first evidence of this is that students in the middle third have college-going rates around 50%, so if all of the differences are due to differences in the proportion of marginal students, one might expect the largest college-going impacts for this group. The results are inconsistent with this idea because the largest increases are among the bottom third. The tests in appendix C support these observations. That is,

the linear relationships between the absolute value of the marginal effects and the absolute deviation of the group's baseline success rate from 0.5 cannot be distinguished from zero (the p -value is .58), and the estimated slope is well outside the range of simulated slopes under mechanical heterogeneity. This suggests that the observed heterogeneity reflects true heterogeneous effects on the underlying disposition to attend college and is not mechanical.

Looking at college type reveals some interesting patterns. Among the bottom three deciles of the educational advantage distribution (who are least likely to attend any college), attending a more effective school increases 2-year college going by 1.1 percentage points, but among the top three deciles (who are more likely to attend college), attending a more effective school reduces 2-year college going by 1.4 percentage points. The increase in college going overall at all advantage levels indicates that this reduction in 2-year college going for the more advantaged students reflects some shifting from 2-year to 4-year programs. In sum, the increase in college going overall is due to increased 2-year and 4-year college going among the bottom third of the educational advantage distribution and an increase in 4-year college going among the top two-thirds of the educational advantage distribution, driven by both (a) increased college going among those who would not have attended college and (b) shifting from a 2-year college to a 4-year college.

Another economically important result is that the increase in 4-year college going is very similar for those in the top and bottom three deciles even though the base rates are very different (about 15% vs. 66%). This indicates that the increases in college going and those for 4-year institutions are not limited to populations with students on the margin. Indeed, the formal tests fail to reject the null hypothesis of no mechanical heterogeneity at the 5% significance level, and one rejects that the observed heterogeneity is all mechanical—indicating larger effects on underlying dispositions to attend 2-year and 4-year colleges among the least educationally advantaged. Remarkably, attending an effective high school can lead to sizable increases in college going, even among student populations for which that may seem unlikely. That is, among the bottom three deciles, a 1 standard deviation increase in effectiveness increases 4-year college going by about 25%.

If the marginal college enrollees from the less advantaged groups are less well prepared for college, they may be less likely to persist and no more likely to earn a college degree (Jackson 2014). Since roughly half of college attrition occurs in the first year (NSCRS 2012), persistence through the first year is a key predictor of college success. As such, we also examine impacts on college persistence beyond freshman year. There are positive impacts throughout the educational advantage distribution, and one cannot reject equality of impacts through the distribution ($p = .19$). Also, there is no linear association between base rates and the marginal effects—indicating sizable benefits to attending more effective schools, particularly for less educationally advantaged

students. From a policy perspective, the similar effects on college persistence across the distribution are important because they imply that the marginal college goers are equally likely to persist irrespective of educational advantage (suggestive of real long-term gains among all students).

Finally, we examine whether a student had ever had a school-based arrest. Among students in the bottom decile, a 1 standard deviation increase in effectiveness decreases in-school arrests by 2.1 percentage points ($p < .01$), compared with only 0.22 percentage points in the top decile ($p < .05$). Even though there are statistically significant effects even among those at the top of the educational advantage distribution, the marginal effects are much more pronounced for those at the bottom. Given the long-term implications of these school-based arrests, this implies sizable long-term benefits to attending effective schools, particularly for those who are least likely to complete high school. Note that this likely represents a lower bound on the effect on arrests because students who may have dropped out of school will not receive a school-based arrest. While these sizable benefits for the least advantaged students are economically important, one may wonder whether the heterogeneity observed reflects differences in the likelihood of being marginal for an arrest. Our formal tests of this suggest that this is the case (see app. C). For arrests, the underlying base rate strongly predicts the marginal effect ($p < .01$), the predicted effect is close to zero for very high and very low base rate groups, and the estimated slope is well within the range of simulated slopes. The data are consistent with effective schools having similar effects on disposition toward crime, which manifest most strongly among those who are marginal. Note that being arrested has serious lifelong implications, so these benefits are real even if the heterogeneity is driven by this mechanical effect.

The results are inconsistent with the notion that the least advantaged are unable to benefit from attending better schools (or benefit less). On the contrary, the evidence is more supportive of the notion that they benefit more—consistent with findings from oversubscribed charter schools in Boston (e.g., Angrist et al. 2012; Walters 2018; Cohodes, Setren, and Walters 2020). To assuage concerns that these patterns reflect comparisons across school type or broad demographic characteristics, appendix D shows that these patterns hold within gender categories, within ethnic/racial groups, and only among traditional public schools.

C. What Is Missed by Test Scores

Because it is common to use test score value-added as the only measure of school quality, we explore how much of school effectiveness (and the heterogeneity in effectiveness) would be missed by only using test score value-added. Before exploring heterogeneity in what test scores miss, we first explore what test scores miss on average. To this aim, table 4 reports the estimated marginal effect (on average for the full population) of overall effectiveness and that of value-added on the individual dimensions. Social and work hard

are very highly correlated, so we combine these two SED measures into a single survey measure.²¹ On average, a 1 standard deviation increase in effectiveness, SED, and behaviors increases test scores by 8.9%, 10.2%, and 5.7% of a standard deviation, respectively. Looking at longer-run outcomes, a 1 standard deviation increase in effectiveness increases high school graduation by 2.4 percentage points, increases college going (within 2 years of high school completion) by 2.57 percentage points, and decreases the likelihood of having a school-based arrest by 0.8 percentage points. In the third row, we show the estimated impact of a 1 standard deviation increase in test score value-added on these same outcomes. Test score value-added predicts impacts on both short- and long-run outcomes, but for all outcomes (including test scores), the improvement associated with a 1 standard deviation increase in effectiveness is greater than that of a 1 standard deviation increase in test score value-added. For the longer-run outcomes, the marginal impacts of the effectiveness index are between 50% and 225% larger than that for test score value-added alone (table 4).²²

To shed light on how the difference between test score value-added and overall effectiveness varies by educational advantage, in figure 4 we plot the difference between the marginal effect for the overall effectiveness and that for test score value-added (along with the confidence intervals for the difference in marginal effects and the linear relationship between the difference and educational advantage).²³ Note that this is not a plot of the marginal effect of non-test-score dimensions of skills but rather a plot of the marginal effect of the dimensions of school quality that are unrelated to test score effects. This analysis explores whether test score value-added misses key dimensions of school quality in ways that vary by educational advantage.²⁴

1. Looking at Short-Run Ninth-Grade Skill Measures

We first examine ninth-grade test scores in the top left panel of figure 4. The plot of overall school effectiveness unexplained by test score value-added

²¹ We provide analogous entries of table 4 in table A6, where the two survey measures are separated.

²² Table 4 also presents the estimated impacts of value-added on the surveys and behaviors. Remarkably, for surveys, test scores, high school graduation, and college enrollment, the school effectiveness index is more predictive of impacts than any of the individual measures. This indicates that the effectiveness index is a good summary measure of school “effectiveness” for these outcomes. However, the behaviors value-added appears to have more predictive power for behaviors and school-based arrests than overall effectiveness. This indicates that school impacts on behaviors capture some meaningful dimension of school quality that is not fully captured by the index and that is predictive of behaviors and school-based arrests.

²³ Practically, we do this by stacking the two regressions into a single model.

²⁴ We test this formally using a stacked linear regression that assesses whether the difference between the marginal effect for the overall effectiveness and that for test score value-added decreases for higher deciles. This is tested using an interaction between the educational advantage and the difference between the two school quality measures.

Table 4
Average Impacts of School Effectiveness and Value-Added

	Test Scores Ninth Grade (1)	Surveys Ninth Grade (2)	Behaviors Ninth Grade (3)	High School Graduation (4)	School- Based Arrests (5)	Enrolled in Any College within 2 Years (6)	Enrolled in 4-Year College within 2 Years (7)	Enrolled in 2-Year College within 2 Years (8)	Persists in College after 1 Year (9)
School effectiveness index	.0890*** (.0133)	.102*** (.00755)	.0571*** (.0127)	.0241*** (.00430)	-.00804*** (.00271)	.0257*** (.00654)	.0416*** (.00903)	-.00570 (.00456)	.0203*** (.00550)
Socioemotional value-added	.0623*** (.0126)	.0840*** (.00818)	.0276** (.0106)	.0197*** (.00397)	-.00564** (.00235)	.0192*** (.00588)	.0321*** (.00866)	-.00541 (.00396)	.0153*** (.00481)
Test score value-added	.0682*** (.0129)	.0445*** (.00849)	.0267*** (.00830)	.0106*** (.00340)	-.00479*** (.00177)	.0168*** (.00564)	.0252*** (.00713)	-.00200 (.00430)	.0137*** (.00498)
Behavior value-added	.0268** (.0122)	.0475*** (.0153)	.228*** (.0164)	.0116** (.00485)	-.0114*** (.00407)	.0167*** (.00558)	.0263*** (.00654)	-.00527 (.00404)	.0151*** (.00396)
Observations	102,235	120,129	157,628	82,146	82,146	55,564	55,564	55,564	55,564

NOTE.—Each point estimate comes from a separate regression. Results are based on regression of outcomes on a single measure of out-of-sample school impacts (overall effectiveness, test score value-added, socioemotional value-added, or behavior value-added). All models include individual demographic controls (race/ethnicity, free and reduced-price lunch, and gender), eighth-grade lags (math and ELA test scores, survey measures, absences, and discipline), and school-level averages for all of the demographics and lagged measures, as well as year fixed effects. We also include the SES of the student census block proxied by average occupation status and education levels. Missing eighth-grade measures were imputed using seventh-grade measures and demographic characteristics. For the longer-run college outcomes, the sample includes two cohorts of first-time ninth graders in spring 2013 and 2014. For the longer-run high school outcomes, the sample includes three cohorts of first-time ninth graders in spring 2012, 2013, and 2014. For the measures, the sample includes six cohorts of first-time ninth graders between spring 2012 and 2017. Note that sample sizes may differ across outcomes because of some missingness in ninth-grade test scores and surveys. Robust standard errors are in parentheses.

** $p < .05$.

*** $p < .01$.

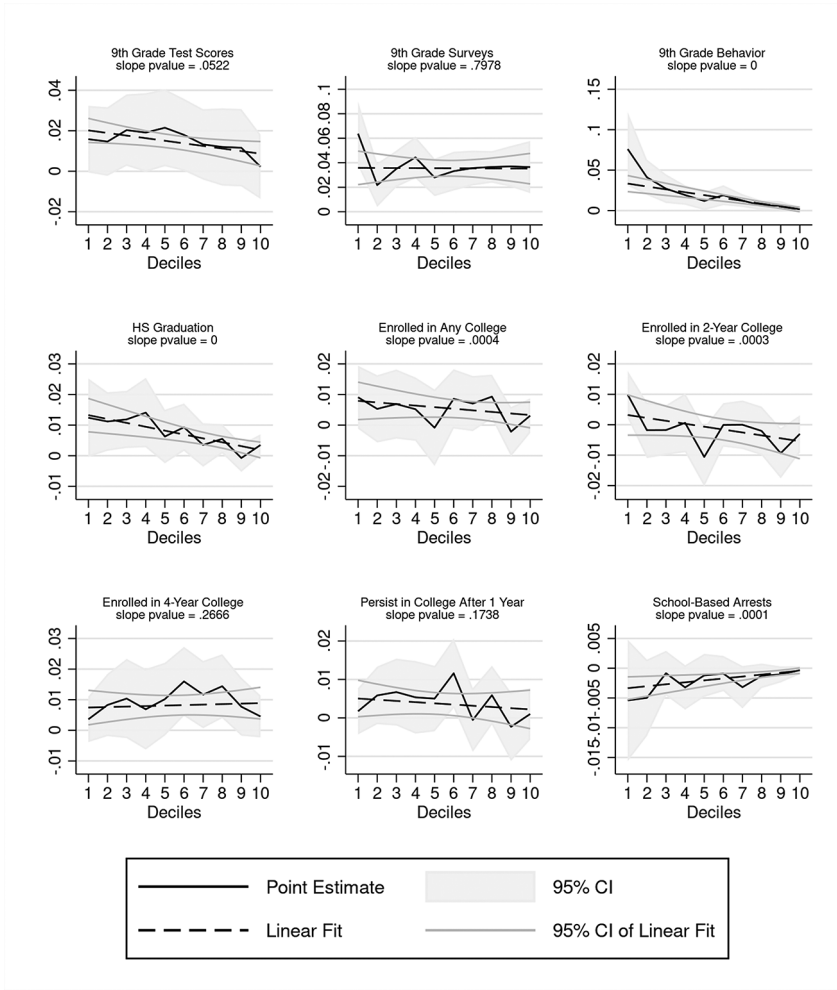


FIG. 4.—Impacts on outcomes unexplained by test score, by educational advantage. Each graph depicts the difference between the marginal impact of a 1 standard deviation increase in the school effectiveness index and in test score value-added. That is, it shows the impact of school effectiveness on each outcome that cannot be explained by test scores. Each regression model controls for the same covariates as in equation (6). Each panel presents the results of 10 separate regressions. The 95% confidence interval (CI) for each point estimate is depicted by the gray area. The dashed black line in each panel depicts the line of best fit for the relationship between deciles of educational advantage and the difference in the marginal effect for the index and test score value-added, including a 95% CI. The reported p -value is for the coefficient on the difference between the interaction between educational advantage and overall effectiveness versus test score value-added in a stacked model. The p -value is associated with the null hypothesis that the difference in effectiveness unexplained by test scores is the same by educational advantage.

against educational advantage has a negative slope ($p < .10$). That is, the test scores of the less advantaged are relatively more responsive to improvements in the dimensions of school quality unrelated to test score value-added than those of more advantaged students. For the most advantaged, the difference between the marginal effect of test scores value-added and the overall index are near zero and statistically indistinguishable from each other, while the marginal effect of the overall index is clearly larger than that of test score value-added for the least advantaged. Looking to the effect on ninth-grade survey-based SED measures, the difference between the marginal effect of the overall index and test score value-added is relatively similar (and positive) through the distribution of advantage. While the slope between the differences in marginal impacts and academic advantage is negative, it is not statistically significantly different from zero ($p = .79$). For behaviors, the slope between the impact unexplained by test score value-added and educational advantage is negative ($p < .01$)—suggesting that the behaviors of the less advantaged are relatively more responsive to improvements in the non-test-score dimensions of school quality than those of more advantaged students.

2. *Looking at Longer-Run Outcomes*

The results for high school graduation are in the middle left panel, and those for college going are in the middle panel. For both of these educational outcomes, the difference between the overall effectiveness effect and that for test score value-added is largest among the least advantaged (about 1 percentage point) and near zero for the most advantaged. For both educational outcomes, the slope between the marginal unexplained effects and educational advantage is negative ($p < .001$). While test score value-added is a reasonable predictor of the educational benefits of attending a more effective school (i.e., effective in multiple dimensions) for educationally advantaged students, this is less true for less advantaged students. Importantly, the potentially heterogeneous effects on college going cannot be due to differences in the likelihood of being marginal across groups (since we find little evidence of this for the overall index). As such, these patterns likely reflect the fact that the educational attainment outcomes of the least advantaged students are relatively more responsive to school impacts on non-test-score dimensions of skills. Finally, we turn to school-based arrests (lower right panel). A plot of the difference between the estimated effects based on overall school effectiveness and test score value-added by educational advantage shows a clear negative relationship ($p < .01$). As with the short-run educational outcomes, the difference between the overall effectiveness effect and that for test score value-added is largest among the least advantaged (about 0.5 percentage points) and near zero for the most advantaged.

To see the importance of the differences in the effects unexplained by test scores by educational advantage, consider the common approach of using average test score value-added to predict effects. While any average measure

of school quality will be somewhat inaccurate for the least advantaged, the extent to which average test score value-added (the most commonly used metric) understates the benefits (as measured by arrests) of attending a more effective school for less advantaged populations is considerable. Looking at high school graduation, ignoring heterogeneity and focusing on test score value-added, the average effect on high school completion of attending a school with 1 standard deviation higher test value-added is 1.06 percentage points (see table 4). Indeed, for students in the top three deciles of the educational advantage distribution, the effect of attending a more effective school overall (using effects on multiple dimensions) is similar (0.78 percentage points), but for students in the bottom three deciles, the overall school effectiveness effect is 3.76 percentage points—more than three times larger than the effect implied by average test score value-added alone. Doing the analogous calculations for college going and school-based arrests reveals that the overall school effectiveness effects on the least advantaged three deciles of students are about 3.7 and 3.1 times larger than the effect implied by average test score value-added alone, respectively. These differences are large, are economically meaningful, and would affect any cost-benefit calculations regarding the benefits of improving schools or any calculation of the distributional effects of improving schools.

In sum, we document a consistent pattern across several outcomes where test value-added misses important dimensions of school quality, particularly for the least advantaged. This pattern may reflect (*a*) general differences in marginal effects across groups due to different groups being differentially marginal for particular outcomes and/or (*b*) less advantaged populations being particularly sensitive to improvements in the non-test-score dimension of school quality. Importantly, the fact that we find similar patterns across all outcomes (including those where we find no evidence that differences in marginal effect are related to differences in being marginal, such as test scores and college going) suggests that the second explanation is at least partially operative. As such, the patterns are consistent with work in psychology suggesting that less advantaged students may enjoy particularly large benefits from interventions that promote SED (e.g., Walton and Cohen 2007; Walton and Cohen 2011; Gray, Hope, and Matthews 2018; Sisk et al. 2018).

D. Policy Implications

1. *Distribution of Effectiveness by Advantage*

Our results indicate that the least advantaged students may benefit the most from attending more effective schools. This implies that investments in high school effectiveness should enhance outcomes for all students while closing outcome gaps between the more and less advantaged. However, to assess the potential for a more targeted approach, we explore the distribution of educational advantage by school effectiveness. At the lowest-performing

schools (bottom 10%), about 15% come from the lowest decile of advantage, while about 3% come from the top decile. Similarly, at the highest-performing schools (top 10%), about 3.7% come from the lowest decile of advantage, while as much as 23% come from the top decile. Consistent with this, on average, those in the top decile attend schools with 1 standard deviation higher effectiveness than those in the bottom decile. However, there is considerable overlap in the quality of schools attended by students across educational advantage groups (see app. E). This indicates that the potential gains to a more equitable distribution of students across schools are economically significant. The patterns also indicate that improving the outcomes of the least educationally advantaged will require targeting schools that disproportionately enroll these students. We consider such a targeted policy below.

2. School Closure Policy

One policy that has been employed is the shutting down of ineffective schools (Brummet 2014) and (importantly) the relocation of affected students to high-performing schools. We run a simple simulation where we remove the lowest-performing 10 schools (8%) and reallocate those students to one of the 10 closest non-low-performing high schools. The static change in value-added is merely the change in the estimated value-added of the attended school (which will mechanically increase for those displaced and be unchanged for other students). As a rough estimate of any indirect effects associated with the reshuffling of peers, we (*a*) estimate the cross-sectional relationship between peer advantage at a school and school value-added and then (*b*) multiply this by the change in peer advantage at the school induced by this policy. See appendix E for additional details.

For the 8% of students who would have attended the least effective schools, the increase in total value-added is 1.78 standard deviations, while the decrease for other students is roughly 0.014 standard deviations (attributed to the inflow of less advantaged peers). This implies sizable benefits for those displaced students. Given the distribution of advantage across schools, for the relocated students high school graduation would increase by about 4.6 percentage points (7.5 percentage points for the least advantaged movers), college going would increase by 4.7 percentage points (7 percentage points for the least advantaged), and school-based arrests would fall by roughly 1.83 percentage points (4.2 percentage points for the least advantaged). For the receiving students, because the spillover effects are small, the changes are very small. We now consider the change in effectiveness for decile groups irrespective of mover and receiver status. The simulated change in effectiveness is very small and positive for the top decile (0.0044 standard deviations), around 0.1 standard deviations for deciles in the middle, and 0.23 standard deviations for the least advantaged decile. That is, even for the most advantaged decile, which has the largest negative peer effect, on average

the total effect is nonnegative—indicating efficiency gains. These gains are even more pronounced because the marginal effects are largest among the most disadvantaged (who experience the largest increases in effectiveness), on average. Because the policy moves only a small fraction of students, the overall effects are modest overall. For the bottom decile, this policy could increase graduation rates by 0.8 percentage points, increase college going by 0.82 percentage points, and reduce arrests by 0.5 percentage points, with smaller effects for the more advantaged groups. We caution readers that our back-of-the-envelope calculations do not capture any disruption effects. However, they do illustrate plausible benefits associated with selectively closing ineffective schools that are disproportionately attended by the least advantaged students.

3. Main Results without Using Survey Measures

Our findings are more broadly applicable if one constructs school effectiveness measures based on readily available data (such as test scores and behaviors) as opposed to surveys (which must be administered). To examine how different our results would be if one excluded school effects on SED (i.e., using test scores and behaviors only), we re-create figure 3 but exclude the school impacts on survey measures (fig. E2). While the pattern of results is very similar, the magnitude (and precision) of the heterogeneity is less than when using all available information (i.e., including the survey measures). For example, while the marginal effects on high school graduation and college going at and above the median advantage are similar with and without SED, for the less advantaged the marginal effects are almost twice as large with SED than without (about 0.02 percentage points on both outcomes without SED, compared with 0.04 percentage points with SED). At the same time, the effects on arrests are quite similar with and without SED. While including the surveys adds additional explanatory power, the results indicate that one can identify schools that will meaningfully impact less advantaged students using readily available administrative data.

V. Conclusions

Whether all students benefit similarly from attending better schools is not well understood. Moreover, the extent to which more or less advantaged students benefit differently from school quality in different dimensions (cognitive vs. socioemotional and behaviors value-added) was unknown. We speak to these issues by examining the effect of attending a more effective school (one that improves a combination of test scores, survey-based SED measures, and behaviors) for more and less advantaged students. Importantly, we do this for a representative set of schools and students, so that our results are more generalizable than existing work. We show that all students benefit from attending effective schools and that the marginal effects

are larger for less advantaged students. While some of the effect heterogeneity in high school completion and arrests may be due to less advantaged groups being marginal for these outcomes, this does not explain larger college-going effects for the least advantaged students (among whom college-going rates are very low). Moreover, insofar as we care about students graduating high school, attending college, and being arrested (irrespective of the mechanism), the documented heterogeneity in benefits is real.

We show that dimensions of school quality unexplained by test score value-added have the largest impacts on the outcomes of less advantaged students—which is, in part, due to less advantaged students benefiting more from non-test-score dimensions of school quality. It is worth noting that our results may be caused by more effective schools focusing inputs on disadvantaged student populations, different students responding differently to the same inputs, or both. While we focus on documenting the existence of this heterogeneity, further work is needed to uncover why—which will, in turn, better inform policy. Our findings reinforce the importance of accounting for soft skills while also accounting for effect heterogeneity. Our results suggest that using test-based measures of school quality alone and ignoring effect heterogeneity may dramatically understate the benefits of attending better schools for those students who may need access to better schools the most.

References

- Agrawal, Mohit, Joseph G. Altonji, and Richard K. Mansfield. 2019. Quantifying family, school, and location effects in the presence of complementarities and sorting. *Journal of Labor Economics* 37:S11–S83.
- Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy* 113:151–84.
- Altonji, Joseph G., and Richard K. Mansfield. 2018. Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects. *American Economic Review* 108:2902–46.
- Angrist, Joshua A., Peter Hull, Parag A. Pathak, and Christopher R. Walters. 2020. Simple and credible value-added estimation using centralized school assignment. NBER Working Paper no. 28241, National Bureau of Economic Research, Cambridge, MA.
- . 2022. Race and the mismeasure of school quality race and the mismeasure of school quality. Working paper.
- Angrist, Joshua D., Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak, and Christopher R. Walters. 2016. Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics* 34:275–318.

- Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. Who benefits from KIPP? *Journal of Policy Analysis and Management* 31:837–60.
- Barrow, Lisa, Lauren Sartain, and Marisa de la Torre. 2020. Increasing access to selective high schools through place-based affirmative action: Unintended consequences. *American Economic Journal: Applied Economics* 12:135–63.
- Beuermann, Diether W., and C. Kirabo Jackson. 2022. The short- and long-run effects of attending the schools that parents prefer. *Journal of Human Resources* 57:725–46.
- Beuermann, Diether W., C. Kirabo Jackson, Laia Navarro-Sola, and Francisco Pardo. 2023. What is a good school, and can parents tell? Evidence on the multidimensionality of school output. *Review of Economic Studies* 90:65–101.
- Brady, Shannon T., Geoffrey L. Cohen, Shoshana N. Jarvis, and Gregory M. Walton. 2020. A brief social-belonging intervention in college improves adult outcomes for black Americans. *Science Advances* 6:eaay3689.
- Bruhn, Jesse. 2020. The consequences of sorting for understanding school quality. Working paper.
- Brummet, Quentin. 2014. The effect of school closings on student achievement. *Journal of Public Economics* 119:108–24.
- Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review* 104, no. 9:2593–632.
- Claro, Susana, David Paunesku, and Carol S. Dweck. 2016. Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences of the USA* 113:8664–68.
- Cohodes, Sarah R., Elizabeth M. Setren, and Christopher R. Walters. 2020. Can successful schools replicate? Scaling up Boston's charter school sector. *American Economic Journal: Economic Policy* 13, no. 1:138–67.
- Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* 78:883–931.
- Davis, Matthew, and Blake Heller. 2019. No excuses charter schools and college enrollment: New evidence from a high school network in Chicago. *Education Finance and Policy* 14:414–40.
- Deming, David J. 2011. Better schools, less crime? *Quarterly Journal of Economics* 126, no. 4:2063–115.
- Deming, David J., Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. 2014. School choice, school quality, and postsecondary attainment. *American Economic Review* 104, no. 3:991–1013.
- Dustan, Andrew, Alain de Janvry, and Elisabeth Sadoulet. 2017. Flourish or fail: The risky reward of elite high school admission in Mexico City. *Journal of Human Resources* 52:756–99.

- Gray, De Leon L., Elan C. Hope, and Jamaal S. Matthews. 2018. Black and belonging at school: A case for interpersonal, instructional, and institutional opportunity structures. *Educational Psychologist* 53:97–113.
- Heckman, James J., and Tim Kautz. 2012. Hard evidence on soft skills. *Labour Economics* 19:451–64.
- Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review* 103:2052–86.
- Hing, Geoff, and Smith Richards Jenniver. 2019. Chicago school choice in charts. *Chicago Tribune*.
- Hoxby, Caroline M., and Sonali Murarka. 2009. Charter schools in New York City: Who enrolls and how they affect their students' achievement. NBER Working Paper no. 14852, National Bureau of Economic Research, Cambridge, MA.
- Jackson, C. Kirabo. 2010. Do students benefit from attending better schools? Evidence from rule-based student assignments in Trinidad and Tobago. *Economic Journal* 120, no. 549:1399–429.
- . 2013. Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics* 95:1096–116.
- . 2014. Do college-preparatory programs improve long-term outcomes? *Economic Inquiry* 52:72–99.
- . 2018. What do test scores miss? The importance of teacher effects on non-test score outcomes. *Journal of Political Economy* 126:2072–107.
- Jackson, C. Kirabo, Shanette C. Porter, John Q. Easton, Alyssa Blanchard, and Sebastián Kiguel. 2020. School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights* 2:491–508.
- Jacob, Brian A. 2002. Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review* 21:589–98.
- Jolliffe, I. T. 2002. *Principal component analysis*. 2nd ed. New York: Springer.
- Kaba, Mariame, and Frank Edwards. 2012. Policing Chicago Public Schools: A gateway to the school-to-prison pipeline. Technical report.
- Kane, Thomas, and Douglas Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. Working paper.
- Kumar, Naveen. 2020. Public school quality and student outcomes: Evidence from model schools in India. Working paper.
- Liu, Airan. 2020. Non-cognitive skills and the growing achievement gap. *Research in Social Stratification and Mobility* 69:100546.
- Liu, Jing, and Susanna Loeb. 2021. Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources* 56, no. 2:343–79.

- Murphy, Mary C., Maithreyi Gopalan, Evelyn R. Carter, Katherine T. U. Emerson, Bette L. Bottoms, and Gregory M. Walton. 2020. A customized belonging intervention improves retention of socially disadvantaged students at a broad-access university. *Science Advances* 6:eaba4677.
- NSCRS (National Student Clearinghouse Research Center). 2012. Fall 2012 yearly success and progress rates. Herndon, VA: NSCRS.
- Oosterbeek, Hessel, Nienke Ruijs, and Inge de Wolf. 2020. Using admission lotteries to estimate heterogeneous effects of elite schools. Working paper.
- Oster, Emily. 2019. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics* 37:187–204.
- Petek, Nathan, and Nolan G. Pope. 2020. The multidimensional impact of teachers on students. Working paper.
- Sass, Tim R., Ron W. Zimmer, Brian P. Gill, and T. Kevin Booker. 2016. Charter high schools' effects on long-term attainment and earnings. *Journal of Policy Analysis and Management* 35, no. 3:683–706.
- Shi, Ying. 2020. Who benefits from selective education? Evidence from elite boarding school admissions. *Economics of Education Review* 74:101907.
- Sisk, Victoria F., Alexander P. Burgoyne, Jingze Sun, Jennifer L. Butler, and Brooke N. Macnamara. 2018. To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science* 29:549–71.
- Spearman, Carl. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15:72–101.
- Walters, Christopher R. 2018. The demand for effective charter schools. *Journal of Political Economy* 126:2179–223.
- Walton, Gregory M., and Geoffrey L. Cohen. 2007. A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology* 92:82–96.
- . 2011. A brief social-belonging intervention improves academic and health outcomes of minority students. *Science* 331:1447–51.