NBER WORKING PAPER SERIES

WHO BENEFITS FROM ATTENDING EFFECTIVE SCHOOLS? EXAMINING HETEROGENEITY IN HIGH SCHOOL IMPACTS

C. Kirabo Jackson
Shanette C. Porter
John Q. Easton
Sebastián Kiguel

Working Paper 28194
http://www.nber.org/papers/w28194

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2020, Revised April 2022

The authors thank the staff at Chicago Public Schools, particularly the Office of Social and Emotional Learning, and the University of Chicago 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 and data management was facilitated by their archivist, Todd Rosenkranz. The authors acknowledge funding for this research from the Bill Melinda Gates Foundation. The content is solely the responsibility of the authors. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Who Benefits From Attending Effective Schools? Examining Heterogeneity in High School Impacts
C. Kirabo Jackson, Shanette C. Porter, John Q. Easton, and Sebastián Kiguel
NBER Working Paper No. 28194
December 2020, Revised April 2022
JEL No. H0,I20,J0

ABSTRACT

We estimate the longer-run effects of attending an effective high school (one that improves a combination of test scores, survey measures of socio-emotional development and behaviours in 9th grade) for students who are more versus less educationally advantaged (i.e., likely to attain more years of education based on 8th-grade characteristics). 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. This heterogeneity is not solely due to less-advantaged groups being marginal for particular outcomes. Commonly used test-score value-added understates the long-run importance of effective schools particularly for less-advantaged populations. Patterns suggest this may, in part, reflect less-advantaged students being relatively more responsive to non-test-score dimensions of school quality.

C. Kirabo Jackson
Northwestern University
School of Education and Social Policy
Annenberg Hall, #204
2120 Campus Dr.
Evanston, IL 60208
and NBER
kirabo-jackson@northwestern.edu

Shanette C. Porter
Mindset Scholars Network
1201 Connecticut Ave. NW
Suite 300
Washington, DC 20036
shanette@gmail.com

John Q. Easton
UChicago Consortium on School Research
University of Chicago
1313 E. 60th St.
Chicago, IL 60637
jqeaston@uchicago.edu

Sebastián Kiguel
Northwestern University
School of Education and Social Policy
2120 Campus Drive
Evanston, IL 60208
skiguel@u.northwestern.edu

I Introduction

A growing body of research in the social sciences finds that schools have important causal effects on students' longer-term outcomes. For example, some charter schools increase college-going (Angrist et al. 2016; Sass et al. 2016), more selective schools improve educational attainment, wages, and health (Jackson 2010; Beuermann and Jackson 2020), and winning a school-choice lottery may increase college-going for girls and reduce interaction with law enforcement among certain boys (Deming et al. 2014; Deming 2011). However, questions remain about whether the benefits of attending better schools differ for better or worse-prepared students. We seek to understand if "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 8^{th}-grade characteristics) students.

In principle, the least educationally advantaged students may benefit most from effective schools because they may have more room for improvement. On the other hand, if "*skills beget skills*" (Cunha et al., 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 schools that rely on admission lotteries or admission tests.[1] As such, existing studies may not generalize to a broad set of traditional schools. Moreover, because these studies rely on comparisons among *applicants* to these special oversubscribed schools (who may differ from typical students), patterns in these studies may differ from those in the broader student population (Bruhn, 2020). That is, disadvantaged students who apply to elite schools or charter school are unlikely to be representative of the typical disadvantaged student (Hoxby and Murarka, 2009). 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 (or non-cognitive skills) may explain attainment gaps by gender (Jacob, 2002) and socio-economic status (Liu 2020; Claro et al. 2016). Moreover, experimental studies in psychology find that (a) students from low-income families or who are academically lower-achieving may benefit from mindset interventions (Sisk et al., 2018), and (b) interventions that promote a sense of belonging are beneficial for the educational outcomes of minoritized (including

[1] Angrist et al. (2012), Cohodes et al. (2020) and Walters (2018) finds 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 schools, Oosterbeek et al. (2020), Barrow et al. (2020), and Dustan et al. (2017) find negative effects of attending elite schools in Amsterdam, Chicago and Mexico City, while Shi (2020) finds the opposite in North Carolina.

Black and Latinx) youth (Gray et al. 2018; Walton and Cohen 2007; Walton and Cohen 2011; Murphy et al. 2020; Brady 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. While Jackson et al. (2020) show that school effects on SED capture important dimensions of school quality *on average*, whether this is especially true for disadvantaged or minority populations is an open question. To shed light on this issue, we identify school effectiveness explained by test score value-added only versus value-added in other dimensions (i.e., socio-emotional development and behaviours), and examine these different effects by educational advantage.

We leverage detailed data from Chicago Public Schools obtained from the UChicago Consortium on School Research. These data link K12 students to high schools and colleges along with test scores, administrative records, and self-reported survey measures of SED over time. Our project entails categorizing students as academically advantaged or not and then estimating the impacts of attending effective schools on these students. This involves three key steps: (1) First we categorize students. We use student behaviours, survey measures, and test scores in 8^{th} grade to predict their educational outcomes years later (dropout, high school graduation, enroll in 2-year college, enroll in 4-year college). We then use this model to create a latent educational advantaged index for each student. (2) Next we identify schools' causal 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 socio-emotional survey measures in 9^{th} grade. We build on Jackson et al. (2020) by *also* estimating school impacts on behaviours (Heckman and Kautz 2012) and combining effects across all outcomes to create an overall school effectiveness index. Providing guidance to policymakers, we describe the characteristics of overall effective schools. We show consistency of our estimates with existing studies, and validate our estimates as reflecting causal impacts. (3) 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. We also explore differences for schools that improve test scores versus other dimensions (i.e., behaviours and survey measures of SED).

First we describe the different indexes. (1) The educational advantage index differentiates between groups of students who are more or less likely to graduate high school, enroll in college, and attend a 4-year college. Students who are low in this index are more likely to have low 8^{th} grade test scores, low socio-emotional measures, and more absences and disciplinary incidents than those who are high on the index. Student low on this index are also more likely to come from low-income homes, and be male and Black – student populations that are hypothesized to benefit the most from socio-emotional interventions. (2) Turning to our measure of school effectiveness, our *overall* school effectiveness index is a better predictor of school effects on longer-run outcomes (i.e., educational attainment and crime) than test score value-added alone. The effectiveness index is weakly related to school demographics and student teacher ratios, and is strongly correlated with college going rates. Consistent with prior research on Chicago schools, selective enrollment schools are somewhat more effective than traditional public schools (consistent with a positive college-going effect found in Barrow et al. (2020) using a regression-discontinuity design)[2], Noble network charter schools are much more effective than other schools (as in Davis and Heller (2019) using lottery assignment), and more typical (i.e., non-Noble) charter schools are similarly effective as traditional public schools (as found more generally in the U.S. Cohodes (2018)) – providing context for, and validating, our index. We also provide evidence that this index captures causal effects.

Looking at heterogeneous effects (our primary focus), *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 short-run outcomes, using formal tests, one cannot reject that the marginal impact of attending a more effective school is the same by educational advantage.[3] That is, the data suggest that attending a more effective schools has roughly the same effect on the short-run skill measures of more and less advantaged students. Looking at the longer-run outcomes, less educationally advantaged students experience larger marginal effects. Specifically, for those in the bottom decile of the distribution attending a school at the 85^{th} percentile of the effectiveness distribution versus one at the median is associated with a 4.4pp increase in high school graduation, a 3.7pp increase in college-going, and a 2.2pp reduction in being arrested – all statistically significant at the 1 percent level. The corresponding estimates for those in the top decile is a 0.9pp increase in high school graduation, a 1.9pp increase in college-going, and a 0.27pp reduction in being arrested. These patterns are similar within ethnic groups and by gender so are not driven by comparisons across broad demographic groups. We test for whether the pattern of effects reflect disadvantaged students being more likely to be marginal for certain outcomes. We find evidence of this for some outcomes but not all – suggesting that the heterogeneity is not entirely mechanical.

Next we examine mechanisms. Looking at college type, all students at more-effective schools are more likely to attend *some* college. However, attending a more effective school increases 2-year and 4-year college going for those at the bottom of the distribution, but shifts students away from 2-year toward 4-year colleges in the middle and top of the distribution. 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

[2] See column 7 of Table 4 in Barrow et al. (2020).

[3] While we do document much larger marginal effects on the behaviours (attendance and discipline) for those at the bottom of the educational advantage distribution, we find that this heterogeneity can be explained by the fact that these behaviours are relatively rare events for the most academically oriented.

suggest that this is due, *in part*, to less academically advantaged students being relatively more responsive to school impacts on SED or soft skills (consistent with interventions in psychology).

The marginal impacts *suggest* that the least advantaged students may gain the most from attending the most effective schools. However, the most advantaged students are most likely to attend highly effective schools. Our estimates indicate that if the least advantaged students (bottom decile) attended the same schools as the most advantaged (top decile), they would be about 4.4pp more likely to graduate high school, 3.7pp more likely to attend college, and about 2.2pp less likely to have a school-based arrest. While differences in school effectiveness do not account for most of the differences in outcomes across students, the potential gains to a more equitable distribution of students across schools are economically meaningful.

By examining impacts for all schools in a district (as opposed to a handful of elite or charter schools) we contribute to the broader school quality literature. We demonstrate that across all public schools in a large district, all students benefit from attending more effective schools. We show sizable increases in college going even among groups with very low college-going rates – reinforcing the policy importance of access to effective schools for disadvantaged students. We also contribute to this literature by moving beyond a test-score measure of effectiveness. By incorporating both psychometrically-sound survey-based measures and behaviour-based measure of soft skills we provide new evidence on different dimensions of school quality captured by each – moving beyond a simple test score vs non-test score paradigm. Moreover, we show how students with varying levels of educational advantage benefit from schools that raise cognitive skills versus socio-emotional skills and behaviours. Finally, we show how test-score measures of school quality may understate the benefits of effective schools – particularly for disadvantaged students.

The remainder of the paper proceeds as follows: Section II described the data used, Section III details the methods we use to categorize students and to measure school effectiveness. Section IV validates our methodology as representing causal impacts. The results are presented in Section V, and Section VI concludes.

II Data

As in Jackson et al. (2020), we use administrative data from Chicago Public Schools (CPS) obtained from the UChicago Consortium on School Research. CPS is a large urban school district with 133 general education public (neighborhood /charter/ vocational/ magnet) high schools serving largely ethnic minority students (41% Black and 45% Latinx) and economically disadvantaged students (86%).[4] The full data-set includes cohorts of 9^{th}-graders who attended one of these schools between 2011 and 2017 (n=157,081). We only include first-time 9^{th} graders to remove sample se-

[4]CPS offers more than 160 programs for students in 9^{th} through 12^{th} grade (link). Our count only includes the general education program offered in neighborhood, charter, vocational and magnet high schools.

lection biases due to grade repetition. For high school graduation and school-based arrests we focus on cohorts of 9^{th} graders between 2011 and 2015 (n=82,146), and for college outcomes we focus on cohorts of 9^{th} graders between 2011 and 2014 (n=55,560) because these students are old enough to have attended college. The data and sample are summarized in Table 1, and discussed below.

Survey Measures: Some of our key variables are survey measures of social-emotional development (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 2008-09, and then every year from 2010-11 onward. These survey items are not part of Chicago's accountability system and response rates were high (78%). However, nonresponse was higher for low-achievers (Appendix Table A1). Note that our analysis of impacts on longer-run outcomes is based on all students irrespective of survey completion. Each survey measure was comprised 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 Appendix Table A2). Following Jackson et al. (2020), we combine the interpersonal-related questions into a **Social Index**[5] and the work-related questions into a **Work Hard Index**.[6] To create each index we standardize each construct, compute the average of the included measures, and then standardize the index to be mean zero and unit variance.

Behavior Measures: Motivated by work showing that impacts on behaviours measure skills not well captured by test score impacts (e.g., Jackson (2018); Liu and Loeb (2019), Heckman et al. (2013), Petek and Pope (2020)), 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, the number of severe disciplinary incidents (eligible for suspension), and the number of days a student is suspended, in each grade. In the analytic sample, the average 9^{th} grader is absent 15.12 days and suspended 0.82 days. Approximately 7.8% of these are involved in a severe disciplinary incident. We summarize these three measures in 9^{th} grade using a **Behaviours Index**. This index is the average of standardized days absent, days

[5]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** includes: 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: 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.

[6]Three survey measures capture students' orientation toward hard work. These are Academic Effort, the perseverance facet of Grit, and Academic Engagement. **Academic Effort** includes: 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: I finish whatever I begin. I am a hard worker. I continue steadily towards my goals. I don't give up easily. **Academic Engagement** includes: 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.

suspended, and severe disciplinary incidents in 9^{th} grade. We standardize the summary measure to be mean zero and unit variance.

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. For each student we average the standardized math and English scores, and then standardize (i.e., make it mean zero with unit variance) this average to create a **Test Score Index**.

Long-Run Outcomes: A key longer-run outcome is having a school-related arrest (among those old enough to have graduated high school). These are arrests for activities conducted on school grounds, during off-campus school activities, or due to a referral by a school official. During our sample period, 3.8 percent of first time 9^{th} graders had a school-based arrest, 5.3 percent of males, and 7.9 percent of Black males. Roughly 20 percent of juvenile arrests in 2010 were school-based arrests (Kaba and Edwards, 2012), so that 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. About 74 percent of first time 9^{th} 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 two years of expected high school graduation (2011 through 2014 cohorts). About 53 percent of first-time 9^{th} graders enrolled in college. We further divide college enrollment into 2-year and 4-year college. In our sample, 34 and 28 percent of students enroll in a 4-year or 2-year college within 2 years of expected graduation, respectively.

III Methods

Our analysis involves three main steps: (1) First, we calculate an educational advantage score for each student by estimating their predicted educational attainment based on a rich set of covariates using an ordered probit. We place students into deciles from least to most likely to attain more years of education. (2) Second, following Jackson et al. (2020), we identify schools that improve students' SED and test scores in 9^{th} grade. In addition, we estimate school value-added on student behaviors using the same method. We combine school effects on the different 9^{th}-grade measures - which are predictive of students' long-term outcomes - into an index of school effectiveness. (3) Finally, we estimate the effect of attending a more effective school among students of differing educational advantage to assess who benefits from attending better schools. We also explore effects of different value-added dimensions to shed light on whether schools that are better in some dimensions (cognitive, socio-emotional, or behaviours) are better for some students than for others.

III.1 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. To classify students, we exploit the fact that we have a rich set of observable characteristics that may predict educational attainment and also multiple measures of educational attainment. In principle, with a single measure of educational attainment (say high school graduation) one could predict high-school completion based on observed covariates in 8^{th} grade. However, because some characteristics may matter more for higher levels of education (such as 4-year college attendance) it is helpful to model the relationship between these covariates and 4-year college going also. If the underlying educational advantage predicts both high-school completion and college-going (or any other educational attainment level), one can model a student's underlying educational advantage (in a way that will predict multiple educational attainment margins) using a rank-ordered probit.

The basic idea is that some underlying educational advantage, y^*, is a linear function of observable characteristics X so that $y^* = X\pi + \varepsilon$. 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

$$y_i = \begin{cases} \text{No High School} & y_i^* \leq \mu_1 \\ \text{Graduate High School} & \mu_1 > y_i^* \leq \mu_2 \\ \text{Attend a 2-Year College} & \mu_2 > y_i^* \leq \mu_3 \\ \text{Attend a 4-Year College} & y_i^* > \mu_3 \end{cases}$$

The probability of observing outcome $y_i = k$ is then $Pr(y_i = k) = Pr(\mu_{k-1} < X\pi \leq \mu_k)$. 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 $(\hat{\pi}, \hat{\mu}_{k-1}, \hat{\mu}_{k-1}, \hat{\mu}_{k-1})$ 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 8^{th} grade math and ELA), 8^{th} grade survey measures, and lagged behaviors.[7] We also include demographics (lunch status, race, gender, and interactions between race and gender). Once the parameter estimates have been estimated, we take the fitted values of latent variable, $X\hat{\pi}$, 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.* As such, each student's predicted educational advantage index is based on the relationship between covariates and educational attainment in *other* cohorts. However, to show the relationship between the advantage

[7] 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 8th grade and an indicator for having any severe disciplinary incidents in 7th or 8th grade.

index and the observable covariates we present the coefficient estimates from the ordered probit model for the full sample in Appendix Table A3.

Differences in Incoming Attributes by Educational Advantage

To shed light on how the attributes of students with high and low educational advantage differ, we present summary statistics for the top and bottom deciles of the educational advantage distribution in the middle and right panels of Table 1. This categorization captures important differences between students, both in terms of demographics and achievement. For example, the top decile contains almost three times more females than the bottom decile (69.9% versus 23.8%), about 8 times fewer students in special education (5.5% versus 45.7%), and less than half the share of students who qualify for free lunch (44.6% versus 94.9%). The top decile has more white students than the bottom decile (22.9% versus 3.8%), more Asian students (18.8% versus 0.13%), but with lower shares of Latinx students (33.1% versus 43%) and Black students (24.1% versus 52.4%). Regarding academic achievement, students in the lowest decile have 8^{th} and 9^{th}-grade test scores more than two standard deviations below those in the top decile. Students in the top decile also have fewer absences (5.6 compared to 34.2 days) and days suspended (0.06 vs. 2.97 days), and are involved in fewer severe incidents (0.007 vs 0.29), relative to the lowest decile in 9^{th}-grade.

Differences in Outcomes by Educational Advantage

To illustrate the differences in our main longer-run outcomes by the latent educational advantage index, we compute the average of our key outcomes for by each percentile of the index. This is presented graphically in Figure 1. This figure highlights a few important facts. First, at the bottom of the index (the bottom 20 percent), even though about 40 percent of students graduate from high school, few (about 17 percent) go to any college, and even fewer (8 percent) attend a 4-year college. Indeed, at the very bottom decile, under 6 percent attend a 4-year college. In the middle of the distribution (between the 40^{th} and 60^{th} percentiles), the high school graduation rate is about 75%, the college-going rate is about 50% and both the 4-year and 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 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. Indeed at the very top of the educational advantage distribution, the 2-year college rate declines with educational advantage. Even though the educational advantage index is predicted based on educational attainment, we also report the school-based arrest rate by educational advantage. School-based arrests are largely concentrated among students with very low educational advantage. For the bottom 20% the arrest rate is roughly 8 percent, while for those above the median it is almost zero (0.02%). Indeed, in the very bottom decile, the arrest rate is a sizable 12.5 percent (see Table 1). It is important to note that even though our educational advantage varies by ethnicity and gender, the patterns also hold within groups. Appendix Figure A1 shows that the average outcomes by educational advantage within groups are very similar to those overall so that the heterogeneity analysis is not merely based on comparisons across the broad demographic groups. Indeed, in Section V.4, we show that while there are meaningful differences, the pattern of results are similar within groups as they are overall.

III.2 Classifying Schools

To isolate schools' causal impacts, we use value-added models to estimate schools' impacts on 9^{th}-grade SED, behaviours, and test scores. We then combine the value-added estimates across outcomes to form an overall school effectiveness index.

Identifying School Impacts on SED, Behaviours, and Test Scores

We seek to isolate the causal effects of individual schools on student measure $q \in Q = \{$test scores, work hard, social, behaviours$\}$ by comparing measure q at the end of 9^{th} grade to those of similar students (with the same survey measures, course grades, incoming test scores, discipline, attendance, and demographics, all at the end of 8^{th} grade) at other schools. School j's value-added on measure q reflects how much school j increases measure q between 8^{th} and 9^{th} grade relative to the changes observed for similar students (based on all the attributes above) who attended different schools. We model the 9^{th} grade measure q of student i who attends school j with observable characteristics Z_{ijt} in year t as (1) below.

$$q_{ijt} = \underbrace{\beta_q Z_{ijt}}_{\text{Effect of Observables}} + \underbrace{\tau_{t,q}}_{\text{Cohort Fixed Effect}} + \underbrace{\alpha_{j,q} + \varepsilon_{ijt,q}}_{\text{Combined error } \upsilon_{ijt,q}} \tag{1}$$

Z_{ijt} includes lagged measures (i.e., 8^{th} grade test scores, surveys, behaviours), gender, ethnicity, free-lunch status, and the socio-economic status of the student's census block.[8] $\tau_{t,q}$ is a cohort fixed effect. To account for correlation between schools and covariates, we follow Chetty et al. (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.[9] Where $\upsilon_{ijt,q} = \alpha_{j,q} + \varepsilon_{ijt,q}$, 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, $\hat{\theta}^{VA}_{jt,q}$. Formally, where N_{jt} is the number of students attending school j in year t,

$$\hat{\theta}^{VA}_{jt,q} = \sum_{i \in jt}^{jt} (u_{ijt,q})/N_{jt} \tag{2}$$

[8]The census block SES measure is the average of occupation status and education levels in the block.
[9]Results are similar using models that exclude $\alpha_{j,q}$ and include school-level averages of the individual covariates.

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 et al. (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.[10] Our leave-year-out predictor for measure q in year t is (3) where the vector of weights $\hat{\psi}_q = (\hat{\psi}_{t-l,q}, ..., \hat{\psi}_{t-1,q}, \hat{\psi}_{t+1,q}, .., \hat{\psi}_{t+l,q})'$ are selected to minimize mean squared forecast errors.

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

A school's predicted value-added on measure q is our best prediction *based on other years* of how much that school will increase measure q between 8^{th} and 9^{th} 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.

Correlations Across Effects on Different Measures

Each value-added measure may represent impacts on a different dimension or may reflect the same underlying school quality. To assess this, we correlate school impacts across these four measures. **For these cross-sectional correlations only** we only need one observation per school. Accordingly, our value-added measure for each school in all years, $\hat{\theta}^{VA}_{j,q}$, uses the full sample and is given by (4) below, where N_j is the number of students assigned to school j across all years.

$$\hat{\theta}^{VA}_{j,q} = \sum_{i \in j}^{j} (u_{ijt,q})/N_j \equiv \underbrace{\theta^{VA}_{j,q}}_{\text{Real value-added on outcome q}} + \underbrace{\zeta_{j,q}}_{\text{Estimation error}} \tag{4}$$

Where q_1 and q_2 connote different outcomes, we report the raw pairwise correlations between the value added for different outcomes (i.e. Corr($\hat{\theta}^{VA}_{j,q_1}, \hat{\theta}^{VA}_{j,q_2}$) in the middle panel of Table 2). Because each of these value-addeds 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. (2022) and implement a split-sample approach that uncovers the real correlation between effect across outcomes under the assumption that estimation errors are unrelated over time both within and across outcomes (as in Kane and Staiger (2008) and Jackson (2013)).[11]

[10] 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).

[11] 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 ($\hat{\theta}^{VA}_{j,,even,q_1} = \theta^{VA}_{j,q_1} + \zeta_{j,even,q_1}$) and those only using the odd years for the other outcome ($\hat{\theta}^{VA}_{j,odd,q_2} = \theta^{VA}_{j,q_2} + \zeta_{j,odd,q_2}$) would not be biased by correlated errors across outcomes because

We report these clean dissattenuated correlations in the bottom panel of Table 2. The general patterns of these clean dissattenuated correlations are broadly similar to those of the raw correlations. Test score value-added is strongly related to the social dimension of SED value-added ($\rho = 0.82$), and has moderate correlations with the work hard dimension of SED value-added ($\rho = 0.313$), and the behaviours value-added ($\rho = 0.435$). This indicates that there may be some underlying dimension of school quality that is associated with higher value-added in all these dimensions. Another interesting pattern is that behaviours value-added are more strongly related to test score value-added than the survey based SED value-added – suggesting that surveys and behaviours may measure different dimensions of socio-emotions skills. Consistent with this notion, the behaviours value added is moderately correlated with the social dimension of SED value-added ($\rho = 0.285$), and largely unrelated to the work hard dimension of SED value-added ($\rho = 0.011$). While, the relatively low correlations among the survey- and behaviour-based measures of socio-emotional skills is an important finding, we leave exploration into *why* to future work.

Creating an Overall School Effectiveness Index

To further understand correlation patterns in the data, we conduct factor analysis of the school effects (Table A5). The model finds that a single underlying factor explains almost all the common variation in these value-addeds. This single factor is positively related to all the value-addeds indicating that it is related to the schools' quality across all dimensions. As such, we combine our value-addeds (work hard, social, test-scores, and behaviours) 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 (5) and represents a measure of school impacts on 9^{th} grade measures that is shared across the SED (work hard and social), test score, and behavior dimensions.[12]

$$\hat{\omega}_{jt} = (0.26)\hat{\mu}_{jt,testscores} + (0.26)\hat{\mu}_{jt,workhard} + (0.34)\hat{\mu}_{jt,social} + (0.14)\hat{\mu}_{jt,behaviors} \tag{5}$$

The weightings indicate that the overall index value-added is positively related to value-added in all dimensions – capturing the best single summary measure of school quality based on these different

$corr(\zeta_{j,even,q_1}, \zeta_{j,odd,q_2}) = 0$. However, this even-odd cross outcome correlation ($\hat{\rho}_{12}^{even-odd}$) 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, $\hat{\rho}_{12}^{even-odd}$) reflects the true correlation times the square root of the product of the reliability of each outcome ((Spearman, 1904)). In this case, that is $\hat{\rho}_{12}^{even-odd} = \rho_{12}^{even-odd}\sqrt{(R_{even,q_1}R_{odd,q_2})}$, where R_{even,q_1} is the reliability of $\hat{\theta}_{j,even,q_1}^{VA}$ and R_{odd,q_2} is the reliability of $\hat{\theta}_{j,odd,q_2}^{VA}$. 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, 1987). The reliability of each measure can be obtained using the correlation between even and odd year estimates for the same outcome (i.e., $\hat{\rho}_{11}^{even-odd}$ and $\hat{\rho}_{22}^{even-odd}$). With the clean raw correlations ($\hat{\rho}_{12}^{even-odd}$) and the estimated reliability ratios ($\hat{\rho}_{11}^{even-odd}$ and $\hat{\rho}_{22}^{even-odd}$), we compute clean dissattenuated correlation estimates $r_{12} = [\hat{\rho}_{12}^{even-odd}]/(\sqrt{(\hat{\rho}_{11}^{even-odd}\hat{\rho}_{22}^{even-odd}})$.

[12] The weights have been normalized to sum to 1 for ease of interpretation.

value-added measures (in that it represents the maximum variance direction in the data (Jolliffe, 2002). We standardize the overall school quality index to be mean zero, unit variance. As we show in Section IV, the index is generally a better predictor of school impacts on longer-run outcomes than the value-added on the individual measures. However, we explore the impacts of test score value-added versus value-added for other dimensions in Section V.

III.3 Some Measurable Differences Between More and Less Effective Schools

To provide some sense of these schools that are effective based on this metric, we present binned scatterplots of observable school attributes by 20 ventiles of school effectiveness in Figure 2. We also report the correlation between the school attribute and school effectiveness. A few patterns emerge. First, despite a strong correlation between average outcomes and student demographics, among the least and most effective schools there are small differences in the share of white students (0.95 vs 0.93 percent in the top and bottom ventiles, respectively), and those on free and reduced-priced lunch (0.92 vs. 0.90 percent in the top and bottom ventiles, respectively). This echoes results from randomized lotteries in New York and Denver (Angrist et al., 2022). The most effective schools tend to be larger that the least effective schools, and have slightly larger class sizes. Noble charter schools are over-represented among the most effective schools (consistent with lottery-based studies (e.g., Davis and Heller (2019)), and selective enrolment schools are slightly more likely to be among the most effective schools (consistent with Barrow et al. (2020) who find positive effects on college-going using a Regression Discontinuity design).[13] As one might expect, given the weak correlations with incoming demographics, more effective schools are also those that have better high school graduation and college gong rates on average.

III.4 Estimating School Effectiveness Impacts by Educational Advantage

Having detailed how we categorize groups of students and groups of schools, we turn to how we estimate the effect of school effectiveness by educational advantage. To quantify the effect of attending a school with one 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 OLS.

$$Y_{ijt} = \delta\hat{\omega}_{jt} + \beta_1 Z_{ijt} + \tau_t + \varepsilon_{ijt} \tag{6}$$

[13] While Barrow et al. (2020) emphasize the somewhat negative effects of selective enrolment schools for some marginal admits, they *do* find statistically significant effects on college-going (see column 7 of Table 4). Moreover, Angrist et al. (2019), show that their somewhat negative results on certain outcomes may not be due to selective schools *per se*, but difference in the counterfactual schools student may attend.

All variables are as defined above and τ_t is a year fixed-effect. In addition to the controls in (1), we include school-level averages of all individual attributes. Standard errors are clustered at the school level.[14] To flexibly present differences in the marginal impacts by student type, we estimate (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 then formally test the hypothesis that the marginal effects are linearly related to educational advantage.[15] If there is no linear relationship, then the slope between effectiveness and the marginal effect will be zero. To take the estimated impacts of effectiveness as reflecting schools' causal impacts requires that, on average, there are no unobserved differences in the determinants of outcomes between students that attend high- and low-effectiveness schools. We provide several empirical tests suggesting causal effects in Section IV.

IV Validating the Method

Test Score Value-Added Versus the School Effectiveness Index

Before exploring heterogeneous effects, we first present the average impacts. Table 3 reports the coefficient on the educational index in a regression of various outcomes on the index and controls for the full sample. The point estimate is the difference in outcomes associated with attending a school with 1σ higher estimated effectiveness (i.e., going from a school at the median to one at the 85^{th} percentile of the effectiveness distribution). As basis for comparison, we also report the estimated effect of the value-added on the individual dimensions also. However, we focus the discussion on the effect on the overall effectiveness index and test score value-added. We refer to schools with a higher estimated overall school effectiveness index as more effective schools.

The top row shows that more effective schools improve 9^{th}-grade test scores, socio-emotional development in 9^{th} grade (as measured by surveys), and behaviours in 9^{th} grade. Specifically, *on average*, a 1σ increase in effectiveness increases test scores by 8.9 percent of a standard deviation, socio-emotional development by 9.3 percent of a standard deviation, and behaviours by 6.3 percent of a standard deviation. Note that social and work hard are very highly correlated so that we combine these two SED measures into a single survey measure.[16] Not surprisingly, more effective schools also improve longer-run outcomes on average. A 1σ increase in effectiveness increases high school graduation by 2.5 percentage points, college going (within 2 years of high school

[14] Individuals with missing 8^{th} grade surveys or test scores are given imputed values. We regress each survey measure or test score on all observed pre-8^{th} grade covariates. We then obtain predicted 8^{th} grade values based on these regressions. and replace missing values with the predictions. Results are similar with and without imputation.

[15] This can be done by running a regression using all the data and interacting all the variables with indicators for the decile of educational advantages and then testing the significance on the interaction between school effectiveness and a scalar of educational advantage. Alternative one can regress δ_d on the decile d while weighting each estimate by the inverse of its estimated variance, $(1/se_{\delta_d})^2$. Both approaches yield similar results

[16] We provide analogous entries of Table 3 in Appendix Table A4 where the two survey measures are separated.

completion) by 2.6 percentage points, and decreases the likelihood of have a school-based arrest by 0.86 percentage points. All of these estimates are significant at the 1 percent level.

Our use of the index (as opposed to using test score impacts only) is motivated by Jackson et al. (2020) showing that a combination of school impacts on test scores and surveys better predict both short and long-run outcomes than test scores alone, and Jackson (2018) showing that a combination of teacher impacts on test scores and behaviours better predict long-run outcomes than test scores alone. We show this to be the case here also. In the third row, we show the estimated impact of a one standard deviation increase in test-score value-added on these same outcomes. Test score value-added *does* predict impacts on both short- and long-run outcomes, but these impacts are smaller than those based on the effectiveness index *including on test scores themselves*. For all outcomes, the improvement associated with a 1σ increase in effectiveness is greater than that of a 1σ increase in test score value-added. For the longer-run outcomes, the marginal impacts of the effectiveness index are between 50 and 250 percent larger than that for test score value added alone.[17] We shed light on the extent to which test scores understate the benefits from attending effective schools varies by educational advantage in V.3 after presenting patterns for the overall effectiveness index.

IV.1 Testing For Selection

Because students are not randomly assigned to schools, one may worry that our effectiveness index is related to unobserved predictors of outcomes so that our estimates are biased. 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.

Control Function Approach

Altonji and Mansfield (2018) show that when individuals or families sort into treatments, group-level averages of observed individual characteristics can potentially serve as a control function and remove all the across-group variation in both observable *and unobservable* individual characteristics. We show that this is likely satisfied in our setting by controlling for (1) school-level average math scores, (2) school-level average absences, and (3) the share of white students at the school. To assess the plausibility of this, we show that conditional on these three school-level averages, predicted outcomes based on a rich set of individual characteristics are unrelated to our effectiveness index. While our test is similar to that in Altonji et al. (2005) and Oster (2019), our control

[17] Table 3 also presents the estimated impacts of value-added on the surveys, and behaviours. 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 behaviours value-added does appear to have more predictive power for behaviours and school-based arrests than the overall effectiveness index. This indicates that school impacts on behaviors capture some meaningful dimension of school quality that is not fully captured by the index and which is predictive of behaviours and school-based arrests.

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.

We predict each outcome based on a linear regression of that outcome on rich individual attributes available in our data (8^{th}-grade test scores, surveys, and behaviours and Grade Point Average (GPA), gender, race, free-lunch status, special education status, and neighborhood socioeconomic status).[18] Figure A1 shows a binned scatterplot of the predicted outcome against the actual outcomes used in this paper. The predicted outcomes track actual outcomes very well.[19] We then examine if the effectiveness index is correlated with predicted outcomes (i.e., a weighted average of *all* observable student-level characteristics that best predicts the student-level outcomes) in a regression model with only the three aforementioned school-level controls.[20] In all models (see columns 4 and 8 in Table 4), school effectiveness is not significantly related to predicted outcomes (conditional on average 8^{th} grade scores, average 8^{th} absences, and percent white at the school) and the point estimates are small. While this shows no selection on *observables* conditional on these key school-level controls, it validates the use of the control function approach suggested in Altonji and Mansfield (2018), to remove selection on unobservables.

Further Evidence of No Selection on Unobservables

Even though conditioning on school-level averages can control for both selection on observables and unobservables in some settings, and we show that this likely holds in our setting, it is helpful to show additional tests. While most schools have residential attendance zones, in Chicago, almost two-thirds of children attend schools other than their zoned school (Hing and Jenniver, 2019). As such, selection on unobservables can occur within attendance zones or across attendance zones. This would occur due to (1) selection of entire families to neighborhoods and therefore zoned schools, and (2) selection of individual students (even within families) to particular schools outside their residential zoned school. We show that neither form of selection seems to be operative in our setting, which taken together, suggest that our estimates are largely unbiased.

Using Variation *Across* Attendance Boundaries. If our results were driven by individual students selecting to schools outside their residentially-zoned schools, then the effectiveness of the residentially-zoned school would be unrelated to student outcomes. We assess this by constructing

[18] Where $(\hat{Y}_{ijt}|Z)$ is the predicted outcome given all the observed covariates, we estimate the following model by Ordinary Least Squares (OLS).

$$(\hat{Y}_{ijt}|Z) = \delta_p \hat{\omega}_{jt} + \tau_t + \upsilon_{ijt} \quad (7)$$

[19] The R-squared is above 0.2 for surveys, behaviours, and test scores, high-school graduation, any college enrollment, and 4-year college enrollment. Those for 2-year college going and arrests are 0.04 and 0.082, respectively.

[20] Recall that we use out-of-sample estimates and equation (3) controls for individual characteristics *within* schools. As such, the fact that we control for some of these observables *within* schools when estimating value-added out-of-sample, **does not** imply no correlation between value-added *across* schools and observables *in-sample*.

instruments that remove the sorting bias that may exists when individuals chose to attend a school outside their zoned area. We instrument for the effectiveness of the school attended with the effectiveness of the residentially-assigned school. **Because families *almost* always have the same address in our database, this approach is largely based on comparisons across families.** The first stage regression is strong – yielding first stage F-statistics above 300. The two-stage-least-squares (2SLS) regressions are reported columns 2,6 and 10 of Table 4. The OLS estimates are reported as a basis for comparison in columns 1, 5, and 9. For the short run outcomes (test scores, surveys, and behaviours), the 2SLS estimates are all positive, significant at the 1 percent level, and of the same order of magnitude as the OLS estimates. Similarly, for all 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 magnitudes as the OLS estimates. For 3 out of 4 outcomes, the 2SLS estimate is significant at the 5 percent level. These patterns rule out that our main results are driven by selection of individual students to schools outside their residential attendance zone. Because these 2SLS models rely primarily on comparisons across families in different attendance zones, these estimates will only be biased if those families that attend the zoned schools self-select into neighborhoods along unobserved dimensions that are correlated with school effectiveness. To rule out this possibility of bias, we also examine variation *within families*.

Using Variation *Within* Families: If the 2SLS results were driven by selection of families to school assignment zones, then there would be no difference in outcomes among members of the same family (in the same attendance zone) but who attend different schools. We find no evidence of this. To isolate within-family variation, we use a subset of the data in which we can identify siblings. We can identify 19,420 families after 2015 in which more than one sibling is observed in 9^{th} grade.[21] There are 7786 families with multiple 9^{th} graders with observed test score measures and 19,420 with observed behaviour measures. Among these families, roughly half have some variation in school attended. While the effective samples are much smaller than the OLS samples for the within-family models (about 14 percent for behaviours and only 9 percent for test scores), they are large enough for reliable within-family estimates of school impacts on the short-run outcomes. We remove the correlation with potentially confounding fixed family characteristics (i.e., the selection of families to neighborhoods) by comparing students from the same family who attended different schools. This is achieved by adding a family fixed effect to our main model in equation (6). The within-family estimates are presented in columns 3, 7 and 11 of Table 4. For the short run outcomes (test scores, surveys, and behaviours), the within-family estimates are all positive, significant at the 1 percent level, and of the same order of magnitude as the OLS estimates – effectively ruling out

[21] Because we cannot identify *all* siblings prior to 2015, these data are imperfect and incomplete. However, if we are able to find similar effects in this small sub-sample as in the broader sample, it would be compelling evidence that our estimates are not biased by family selection to neighborhoods.

that the 2SLS results were driven by selection of families to residential zones.

For longer-run outcomes measured at the end of high school, there is considerably less data – limiting our ability to reliably estimate within-family models. There are 3,902 families with multiple children old enough to have graduated high school. This yields effective samples much smaller than the OLS samples for the within-family models on these outcomes (about 5 percent for high school graduation and arrests), which may not be reliable. Consistent with this, the standard errors for the within family estimate on these two outcomes are much larger than those for OLS. However, the within-family estimates are both similar to the OLS estimates and significant at the 5 percent level. Unfortunately the effective sample for college-outcomes is less than half than for high school outcomes, so due to lack of sufficient data, we cannot reliably estimate the within-family models for the college outcomes.[22] However, the consistent within-family patterns for the short and medium term outcomes suggest that this would likely also hold true for the college outcomes.

Considering all the Tests Together

If our estimates were biased by selection, one would expect that strong predictors of outcomes would 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 2SLS and within-family results (coupled with the lack of selection on observables conditional on the three school-level variables) suggest that neither is biased. While none of these tests is dispositive in isolation, together they are compelling evidence that our estimated school impacts, and the main results, reflect true causal impacts.

V Results

V.1 Impacts of School Effectiveness on Short-Run Measures: By Advantage

We now consider how effects vary for students with different levels of *ex-ante* educational advantage. We estimate these same regressions for each measure and for each decile. To summarize these ten regressions per outcome, we plot the point estimate and 95% confidence intervals for each estimate in Figure 3. Appendix Table A7 reports the point estimates and standard errors for select deciles. We also plot the linear relationship between the estimated effect and the educational

[22]There are only 1,634 families with multiple children old enough to have enrolled in college. Among these, very few families have variation in school attended and college status – so that the within family models are inconclusive on this sample. To show the implications of the lack of variation in this sub-sample, in Table A6 we show that using the sample of families for which there multiple children old enough to have enrolled in college none of the within-family relationships shown for test scores, surveys, and behaviours persist – indicating that there is not enough variation among this sub-sample for reliable inference for *any* outcome.

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 one standard deviation higher in school effectiveness on students' 9^{th} grade test scores. All students benefit from attending a more effective school. While the point estimates suggest that the marginal benefits are larger for the least educationally advantaged, the p-value associated with the hypothesis that the effect varies linearly with educational advantage is 0.21 – indicating that the impacts of attending a more effective school on 9^{th} grade test scores are similar throughout the educational advantage distribution.

We now turn to the survey measures. The middle panel shows the effect of attending a school one standard deviation higher on the school effectiveness index on students' socio-emotional measures in 9^{th} grade by decile of predicted educational attainment. As with test scores, all students benefit from attending a more effective school, and one cannot reject that the impacts of attending a more effective school on 9^{th} grade survey measures are the same throughout the educational advantage distribution. We report the results for 9^{th} grade behaviours in the right top panel of Figure 3. Unlike the socio-emotional and test score measures, one rejects the hypothesis of the same marginal effect for all advantage groups at the 1 percent level. Effective schools have the strongest effect on the observed behaviors for students in the lower end of the advantage distribution. For a student in the lowest (first) decile, attending a school 1 standard deviation higher in school effectiveness improves the behavior index by 0.19 standard deviations. Meanwhile, for students in the top (tenth) decile, the behavior index only improves by 0.015 standard deviations. One interpretation of this pattern in 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 to 0.293 in the bottom decile) and have a low absence rate (5.6 days compared to 34.2 days in the bottom decile), so that there is relatively little room for improvement. We assess this in Section V.1.1. Using a formal test, we show that the heterogeneous impacts on behaviours is likely due to differences in the likelihood of being "marginal" across groups, rather than reflecting heterogeneous impact on underlying behaviour-related skills *per se*.

V.1.1 Testing for Mechanical Heterogeneity

The basic idea of this test is that for most models of binary outcomes (such as a log, 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., zero or one).[23] Under purely mechanical heterogeneity (i.e., the same underlying change in latent disposition across groups), there will be (a) a negative relation-

[23] We expand upon the logic of this test formally in Appendix B.

ship 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 of from 0.5 of 0.5. We assume a linear relationship, for simplicity, and test for this using the regression in (8) below.

$$|\delta_g| = \alpha + \pi \times (|p_g - 0.5|) + \upsilon_g \tag{8}$$

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 behaviours provides strong evidence of such "mechanical" effects (see Appendix Figure B1). For the likelihood of being chronically absent, having any suspensions, or having any disciplinary infections, one rejects that the slope is zero at the 1 percent level and one cannot reject that the effect is zero for those who are very likely or unlikely to have a success. This result is consistent with the similar effect on the continuous measures (i.e., test scores and surveys) across the educational advantage distribution – suggesting little differential effect on underlying skills in the short run.

V.2 Impacts on Longer-Run Outcomes: By Advantage

Having shown the effects on short-run measures in 9^{th} grade, we now examine similar figures for the longer-run outcomes (the middle and lower panels of Figure 3). Looking at high school graduation, the marginal impacts of school effectiveness are much larger for students at the bottom of the educational advantage index than those at the top. One rejects that the linear relationship between the marginal effect and educational advantage is zero at the 1 percent level. Indeed, for those in the bottom decile, a 1σ increase in effectiveness increases high school completion by 4.4 percentage points (p-value$<$0.01) compared to only 0.9 percentage points (p-value$<$0.10) in the top decile. Relative to each groups' baseline level, this is about a 15 percent increase for those at the bottom of the distribution compared to a 2 percent increase for the top. While the differences in the changes in graduation rates across groups are real and economically meaningful, one may wonder if 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 V.1.1. The results of this test are summarized in Appendix Figure B1, which plots the linear relationships between absolute value of the marginal effects against the absolute deviation of the groups baseline success rate from 0.5. For high school graduation, the p-value on the slope is significant at the 1 percent level and the implied effect is near zero for groups that are very likely or unlikely to have a success – suggesting that the considerable heterogeneity in the effect on high-school graduation by educational advantage *may* be due to differences in baseline success rates across groups.

Next we examine is enrolling in any college (2-year or 4-year) within two 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 0.002). More specifically, for the bottom third, a 1σ increase in effectiveness increases college-going by 4.1 percentage points (p-value$<$0.01) compared to about 1.5 percentage points (p-value$<$0.01) in the top third. Given the large differences in base rates, the differences in relative marginal impacts are sizable. For the bottom third, a 1σ increase in effectiveness increases college-going by about 22 percent compared to about 3 percent in the top third. 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 percent, so that 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; if anything, the largest increases are among the bottom third. The regression test in Appendix Figure B1 supports these observations. That is, the linear relationships between absolute value of the marginal effects and the absolute deviation of the groups baseline success rate from 0.5 cannot be distinguished from zero – the p-value is 0.78. This suggests that the observed heterogeneity reflects heterogeneous effects on the underlying disposition to attend college.

Looking at college type reveals some interesting patterns. The average results in Table 3 suggested no impacts on 2-year college going. However, the heterogeneous impacts provide an explanation for the null result on average. Among the bottom third of the educational advantage distribution (who are least likely to attend *any* college), attending a more effective school increases 2-year college going by roughly 1.4 percentage points (p-value$<$0.05), but among the top two-thirds (who are more likely to attend college), attending a more effective school *reduces* 2-year college going by about 1.4pp. The increase in college-going overall indicates that this reduction in 2-year college going for the advantaged students reflects shifting from 2-year to 4-year programs. Once can see this clearly when looking at 4-year college going. While all groups have increased 4-year college going, the groups with the largest increases in 4-year college going are those with the reductions in 2-year college going. Specifically, for the bottom third, a 1σ increase in effectiveness increases 4-year college-going by about 3.9pp compared to over 5pp for the middle third and 3.1pp for the top third. 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 similar for those in the top third and bottom third even though the base rates are very different (15 versus 66 percent). This indicates that the increases in college going, and those for 4-year institutions are not limited only to populations with students on the margin. Indeed, the formal tests fail to reject the null hypothesis of no "mechanical heterogeneity" – 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 third, a 1σ increase in effectiveness increases 4-year college-going by about 25 percent.

If the marginal college enrollees from the less advantaged groups are less well-prepared for college, they may be less likely to persist in college, and no more likely to earn a college degree (Jackson, 2014). Since most college attrition occurs in the first year, 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-value=0.82). Also, there is no linear association between base rates and the marginal effects – indicating sizable benefits to attending more effective school particularly for less educationally-advantaged students. From a policy perspective, the similar effects on college persistence across the distribution is important because it implies 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. Because this is a relatively rare outcome among students at the top of the educational advantage distribution, one would not expect much effect at the top of the distribution. Indeed, this is what one observes. Among students in the bottom decile, a 1σ increase in effectiveness decreases in-school arrests by 2.2 percentage points (p-value<0.01) compared to only 0.27pp in the top decile (p-value<0.05). 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. It is worth noting 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 real and economically important, one may wonder whether the heterogeneity observed reflects differences in the likelihood of being marginal for an arrest. Our formal test of this suggests that this is the case (see Appendix Figure B1). For arrests, the underlying base rate strongly predicts the marginal effect ($p < 0.01$), and the predicted effect is close to zero for very high and very low base rate groups. As such the data are consistent with effective schools having a similar effects on disposition toward crime, but that this manifests most strongly among those in the bottom deciles (who have baseline rates closer to 0.5).

Despite theories suggesting that the gains should be largest for the least advantaged and others suggesting the opposite, the results show that all students benefit from attending more effective schools. 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), Cohodes et al. (2020) and Walters (2018)).

V.3 Heterogeneity in the Importance of Dimensions Missed by Test Scores

Our effectiveness measure reflects a combination of school impacts on test scores, surveys, and behaviours. One might wonder if measures of school quality that exclude impacts on SED miss important components of school quality particularly for disadvantaged populations. To shed light on this, 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).[24] Note that this *is not* a plot of the marginal effect of non-test score dimensions of skills, but 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.

Looking at Short Run 9^{th} Grade Skill Measures

Looking at the short-run outcomes, the non-test score dimensions of school quality have more explanatory power (above and beyond test score value-added) for the less educationally advantaged. We first examine 9^{th} grade test scores in the top left panel of Figure 4. The plot of overall school effectiveness unexplained by test score value-added against educational advantage has a negative slope (p-value$<$0.01). 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. Indeed, 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 9^{th} 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-value=0.45). For behaviours, the school effects unexplained by

[24] Practically, we do this by stacking the two regressions into a single model.

test-score value-added is appreciably larger for the less-advantaged. The slope between the impact unexplained by test score value-added and educational advantage is negative (p-value<0.01)- suggesting that the behaviours for the less advantaged are *relatively* more responsive to improvements in the non-test score dimensions of school quality than those of more advantaged students. We caution, that because there is less variation in behaviours for the most advantaged, this particular result for behaviours may be somewhat mechanical.

In sum, for all three 9^{th} grade outcomes, the impacts of school effectiveness unexplained by test score value-added is larger for the less advantaged, and the slope is statistically significant for two of them (including test scores for which there is similar variation among more and less advantaged groups). Taken the together, this suggests that the extent to which test score value-added misses key components of school quality *tends to be* greatest for less-advantaged students. Put differently, in terms of the short-run outcomes, the least-advantaged students appear to benefit the most from those components of school quality unmeasured by test score value-added. We now show that this pattern also holds for the longer-run outcomes.

Looking at Longer-Run Outcomes

The results for high school graduation are in the middle left panel. The difference between the overall effectiveness effect and that for test-score value-added is largest among the least advantaged (about 2pp) and near zero for the most advantaged. Indeed, the slope between the marginal unexplained effects and educational advantage is negative (p-value<0.01). As with the short-run measures, while test score value-added is a reasonable predictor of the benefits of attending a more effective school (that is, effective in multiple dimensions) on high-school graduation for educationally advantaged students, test score value-added may be a particularly poor predictor of overall school effects on high school completion for less-advantaged students. To see the importance of this, consider the common approach of using average test score value added to predict effects. Ignoring heterogeneity and focusing on test score value-added, the average effect on high-school completion of attending a school with 1σ higher test value-added is 0.94pp (see Table 3). Indeed, for students in the top third of the educational advantage distribution, the effect of attending a more effective school overall (using effects on multiple dimensions) is similar (0.99pp), but for students in the bottom third, the overall school effectiveness effect is 4pp – more than four times larger that the effect implied by average test score value-added alone. These differences 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.

The pattern is similar for college-going (middle panel). The marginal school-effectiveness effect unexplained by test score value-added is larger for less-advantaged students. For the bottom three deciles, the marginal effect of the index is about 1pp larger than that for test score value-added, while this difference is near zero for the most advantaged. However, these patterns are only suggestive because one fails to reject that the unexplained gap is the same for all groups at the 0.05 significance level. 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, this pattern is likely driven by the fact that the college-going outcomes of the the least-advantaged students are relatively more responsive to school impacts on non-test-score dimensions of skills. As with high-school graduation, average test score value added is much worse predictor of school effects on college-going for less-advantaged students than the more-advantaged. Specifically, the average effect on college going of attending a school with 1σ higher test value-added is 1.5 percentage points. For students in the top third of the educational advantage distribution, the effect of the overall index is similar (about 1.5pp), but for students in the bottom third of the educational advantage distribution, it is about 4.1pp – more than twice as large as that implied by average test score value-added alone.

We now turn to school-based arrests (lower right panel). A plot of the difference between the overall effectiveness effect and test score value-added by education advantage shows a clear negative relationship ($p-value < 0.01$). We caution that this pattern *may* be an artifact of different groups being differentially marginal for arrests. Irrespective of the reasons, given that having an arrest has real economically meaningful implications, the documented heterogeneity has real world and policy implications. To see the importance of accounting for heterogeneity and also school effects on non-test score dimensions, consider the following calculations. The average effect on the likelihood of arrest of attending a school with 1σ higher test value-added is 0.4pp. For students in the top third of the educational advantage distribution, the effect of the overall index is similar (about 0.37 percentage points), but for students in the bottom third of the educational advantage distribution, the effect of attending a 1σ more effective school is about 1.6pp – more than four times as large. The extent to which test score value-added understates the benefits (as measured by arrests) to attending a more effective school for less-advantaged populations is considerable.

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. We are careful to note that this pattern may be due to (a) general differences in marginal effects across groups due to different groups being differentially marginal for particular outcomes, and/or (b) less educationally advantaged populations being particularly sensitive to improvements in the non-test score dimension of school quality. We cannot *rule out* the first explanation. However, 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 partly operative. As such, the patterns are broadly consistent with work in psychology suggesting that less advantaged students may enjoy particularly large benefits from interventions that promote socio-emotional development (Sisk et al. 2018,Gray et al. 2018; Walton and Cohen 2007; Walton and Cohen 2011).

V.4 Differences by Race and Gender and School Type

The summary statistics in Table 1 show that students in the bottom and top of the educational advantage distribution differ along both sex and ethnicity dimensions. As such, one may wonder if these patterns reflect gender or race differences, or if these are broad patterns that exist within demographic groups. To assess this, we implement analogous analyses using students from a particular group (males, females, black, Latinx). Appendix Figure A2 shows that the average outcomes by educational advantage within groups are similar to those overall so that the heterogeneity analysis within groups is meaningful. By and large, the patters of results that we document across all groups exists within groups. As such, our results are not an artifact of making comparisons across sex or ethnic groups. There are, however, some differences we detail below.

In Figure 5, the marginal effects are quite similar for both males and females; the average effects on high-school graduation are slightly larger for males, the effects on college going are slightly larger for females, and the effects on arrests rates of males somewhat larger for males that for females. Importantly, for both males and females, the less educationally-advantaged students experience larger marginal effects from attending a more effective school.

In Figure 6 we show effects for Black and Latinx students separately (other ethnic groups are too small to examine heterogeneous impacts). The arrest outcomes are much more sensitive to school effectiveness for Black students than Latinx students, while the educational attainment effects are particularly pronounced for Latinx students. In particular, among Black students in the bottom decile of the educational advantage distribution, a standard deviation increase in school effectiveness reduces the likelihood of a school-based arrest by about 3 percentage points (p-value$<$0.01), while that for Latinx students is about one percentage point. Looking at educational outcomes, for Latinx students in the bottom of the educational advantage distribution, a standard deviation increase in school effectiveness increases the likelihood of high-school graduation by over 5 percentage points (p-value$<$0.01), and in the middle of the distribution it increases the four-year college going rate by around 9 percentage-points.[25] The analogous numbers for Black students are around 3 percentage points for high school graduation and 1.5 percentage points for four-year college going. For both Black and Latinx students, the less educationally-advantaged students experience larger marginal effects from attending a more effective school. However, this heterogeneity by educational advantage is more pronounced for Latinx students – suggesting that less educationally-advantaged Latinx students may be particularly well-served by access to high-quality schools.

[25]These relatively large college-going effects are consistent with Jackson (2014) finding particularly large college going responses among Latinx student to a college preparatory program in Texas.

Since much evidence of differential school effectiveness is based on small samples of oversubscribed charter schools, one may wonder if our results hold only among traditional public schools. To assess this, we implement the entire analysis looking only at traditional public schools (Appendix Figure A3). The patterns we document are generally similar when restricted only to traditional public schools. This suggests that the patterns we document may generalize to other settings.

V.5 Distribution of Effectiveness by Advantage

Our results indicate that the least educationally advantaged students may benefit the most from attending more effective schools. As such, it is instructive to assess whether school effectiveness is evenly distributed by educational advantage. To this aim, we compute various percentiles of the school effectiveness index for students in each decile of the advantage index. This provides information about the extent of exposure to high-quality schools by educational advantage. We plot the percentiles for the deciles in Figure 7. One takeaway from this figure is that students of all educational advantage levels are exposed to schools that are both high and low on the effectiveness index. Indeed, the differences in school effectiveness within each decile (e.g., comparing the 5^{th} to the 95^{th} percentile of school effectiveness within a given educational advantage decile) are much larger than the differences in the same percentiles of effectiveness across educational advantage (e.g., comparing the 95^{th} percentile of school effectiveness for the top and bottom deciles of educational advantage). However, there *are* economically significant differences across deciles.

Looking across deciles of educational advantage, the the most advantaged are exposed to more effective schools. Indeed, the 95^{th} percentile of school effectiveness for the bottom and top deciles are about 1.4 and 2.4 respectively - a sizable 1σ difference. The estimates in Figure 3 and Appendix Table A7 indicate that a 1σ increase in effectiveness for the bottom decile would increase high school graduation by about 4.4pp, college-going by about 3.7pp, and reduce the likelihood of being arrested by 2.2pp. The differences are similar at the median and somewhat smaller at the 5^{th} percentile of school effectiveness (a difference of about 0.5σ) – generally indicating economically important differences for the more advantaged compared to the less advantaged. While differences in school effectiveness do not account for most of the differences in outcomes across students with differing levels of educational advantage (see Figure 1), the potential gains to a more equitable distribution of students across schools are economically significant.

VI Conclusions

It is known that schools can have meaningful impacts on both short- and longer-run outcomes. However, 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 versus socio-emotional and behaviours value added) is 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 behaviours) 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 is due to less-advantaged groups being marginal for some outcomes, this "mechanical heterogeneity" *does not* explain larger college-going effects for the least advantaged student (among which college going rates are very low). We show that dimensions of school quality unexplained by test-score value-added have the largest impacts for the less-advantaged students– which is, in part, due to less-advantaged students benefiting more from non-test score dimensions of of school quality. Our findings reinforce the importance of accounting for soft skills while also accounting for effect heterogeneity. The patterns we uncover suggest that if one were to use test-based measures of school quality alone and ignore effect heterogeneity, one would dramatically understate the benefits to attending better schools for those students who may need access to better schools the most.

References

Joseph G. Altonji and Richard K. Mansfield. Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects. *American Economic Review*, 108: 2902–46, 10 2018. ISSN 0002-8282. doi: 10.1257/AER.20141708.

Joseph G. Altonji, Todd E. Elder, and Christopher R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113:151–184, 2 2005. ISSN 00223808. doi: 10.1086/426036/0.

Joshua A Angrist, Peter Hull, Parag Pathak, and Christopher R Walters. Race and the mismeasure of school quality race and the mismeasure of school quality *. *Working Paper*, 2022. URL https://www.usnews.com/education/k12/.

Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. Who benefits from kipp? *Journal of Policy Analysis and Management*, 31:837–860, 9 2012. ISSN 02768739. doi: 10.1002/pam.21647. URL http://doi.wiley.com/10.1002/pam.21647.

Joshua D. Angrist, Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak, and Christopher R. Walters. Stand and deliver: Effects of boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, 34:275–318, 4 2016. ISSN 0734306X. doi: 10.1086/683665. URL https://www.journals.uchicago.edu/doi/abs/10.1086/683665.

Joshua D Angrist, Parag A Pathak, and Román Andrés Zárate. Choice and consequence: Assessing mismatch at chicago exam schools choice and consequence: Assessing mismatch at chicago exam schools. 2019. URL http://www.nber.org/papers/w26137.

Lisa Barrow, Lauren Sartain, and Marisa de la Torre. Increasing access to selective high schools through place-based affirmative action: Unintended consequences. *American Economic Journal: Applied Economics*, 2020. ISSN 1945-7782. doi: 10.1257/APP.20170599.

Diether Beuermann, C Kirabo Jackson, Laia Navarro-Sola, Francisco Pardo, and Inter-American Development Bank. What is a good school, and can parents tell? evidence on the multidimensionality of school output. *Review of Economic Studies*, 12 2022. ISSN 1556-5068. doi: 10.3386/W25342. URL https://www.nber.org/papers/w25342.

Diether W. Beuermann and C. Kirabo Jackson. The Short and Long-Run Effects of Attending The Schools that Parents Prefer. *Journal of Human Resources*, pages 1019–10535R1, apr 2020. ISSN 0022-166X. doi: 10.3368/jhr.57.3.1019-10535r1. URL http://jhr.uwpress.org/content/early/2020/04/03/jhr.57.3.1019-10535R1http://jhr.uwpress.org/content/early/2020/04/03/jhr.57.3.1019-10535R1.abstract.

Shannon T. Brady, Geoffrey L. Cohen, Shoshana N. Jarvis, and Gregory M. Walton. A brief social-belonging intervention in college improves adult outcomes for black americans. *Science Advances*, 6:eaay3689, 4 2020. ISSN 23752548. doi: 10.1126/sciadv.aay3689. URL http://advances.sciencemag.org/.

Jesse Bruhn. The consequences of sorting for understanding school quality, 2020.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104(9):2593–2632, sep 2014.

Susana Claro, David Paunesku, and Carol S. Dweck. Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences of the United States of America*, 113:8664–8668, 8 2016. ISSN 10916490. doi: 10.1073/pnas.1608207113. URL www.pnas.org/cgi/doi/10.1073/pnas.1608207113.

Sarah Cohodes. Charter schools and the achievement gap, 2018. URL https://futureofchildren.princeton.edu/news/charter-schools-and-achievement-gap.

Sarah R. Cohodes, Elizabeth M. Setren, and Christopher R. Walters. Can successful schools replicate? scaling up boston039;s charter school sector. *American Economic Journal: Economic Policy*, 2020. ISSN 1945-7731. doi: 10.1257/POL.20190259.

Flavio Cunha, James J. Heckman, and Susanne M. Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78:883–931, 5 2010. ISSN 0012-9682. doi: 10.3982/ecta6551. URL https://onlinelibrary.wiley.com/doi/full/10.3982/ECTA6551https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6551https://onlinelibrary.wiley.com/doi/10.3982/ECTA6551.

Matthew Davis and Blake Heller. No excuses charter schools and college enrollment: New evidence from a high school network in chicago. *Education Finance and Policy*, 14:414–440, 7 2019. ISSN 1557-3060. doi: 10.1162/EDFP_A_00244.

David J. Deming. Better Schools, Less Crime? *. *The Quarterly Journal of Economics*, 126(4):2063–2115, nov 2011.

David J. Deming, Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. School Choice, School Quality, and Postsecondary Attainment, 2014.

Andrew Dustan, Alain de Janvry, and Elisabeth Sadoulet. Flourish or fail: The risky reward of elite high school admission in mexico city. *Journal of Human Resources*, 52:756–799, 6 2017. ISSN 15488004. doi: 10.3368/jhr.52.3.0215-6974R1. URL http://jhr.uwpress.org/content/52/3/756http://jhr.uwpress.org/content/52/3/756.abstract.

De Leon L. Gray, Elan C. Hope, and Jamaal S. Matthews. Black and belonging at school: A case for interpersonal, instructional, and institutional opportunity structures. *Educational Psychologist*, 53:97–113, 4 2018. ISSN 00461520. doi: 10.1080/00461520.2017.1421466. URL https://www.tandfonline.com/doi/abs/10.1080/00461520.2017.1421466.

James Heckman, Rodrigo Pinto, and Peter Savelyev. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103:2052–2086, 10 2013. ISSN 00028282. doi: 10.1257/aer.103.6.2052.

James J. Heckman and Tim Kautz. Hard evidence on soft skills. *Labour Economics*, 19:451–464, 8 2012. ISSN 0927-5371. doi: 10.1016/J.LABECO.2012.05.014.

Geoff Hing and Smith Richards Jenniver. Chicago school choice in charts - Chicago Tribune, 2019. URL https://www.chicagotribune.com/ct-chicago-school-neighborhood-enrollment-charts-20160106-htmlstory.html.

Caroline M Hoxby and Sonali Murarka. Charter schools in new york city: Who enrolls and how they affect their students' achievement. 4 2009. doi: 10.3386/W14852. URL https://www.nber.org/papers/w14852.

C. Kirabo Jackson. Do Students Benefit from Attending Better Schools? Evidence from Rule-based Student Assignments in Trinidad and Tobago*. *The Economic Journal*, 120(549):1399–1429, dec 2010. ISSN 00130133. doi: 10.1111/j.1468-0297.2010.02371.x.

C. Kirabo Jackson. Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics*, 95:1096–1116, 10 2013.

C. Kirabo Jackson. Do college-preparatory programs improve long-term outcomes? *Economic Inquiry*, 52: 72–99, 1 2014. ISSN 1465-7295. doi: 10.1111/ECIN.12040. URL https://onlinelibrary.wiley.com/doi/full/10.1111/ecin.12040https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12040https://onlinelibrary.wiley.com/doi/10.1111/ecin.12040.

C. Kirabo Jackson. What do test scores miss? the importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, 5 2018. doi: 10.3386/w22226.

C. Kirabo Jackson, Shanette C. Porter, John Q. Easton, Alyssa Blanchard, and Sebastián Kiguel. School effects on socio-emotional development, school-based arrests, and educational attainment. *American Economic Review: Insights*, 2020. doi: 10.1257/AERI.20200029.

Brian A. Jacob. Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review*, 21:589–598, 12 2002. ISSN 02727757. doi: 10.1016/S0272-7757(01)00051-6.

I T Jolliffe. *Principal Component Analysis, Second Edition*. Springer, 2002.

Mariame Kaba and Frank Edwards. Policing Chicago Public Schools:A Gateway to the School-to-Prison Pipeline. Technical report, 2012. URL http://cpdincps.com/fullreport.

Thomas Kane and Douglas Staiger. Estimating teacher impacts on student achievement: An experimental evaluation, 12 2008.

Naveen Kumar. Public school quality and student outcomes: Evidence from model schools in india, 2020. URL https://drive.google.com/file/d/1-Va6T6WLZzIi1HOUpTBxROWUoO4takXz/view.

Airan Liu. Non-Cognitive skills and the growing achievement Gap. *Research in Social Stratification and Mobility*, page 100546, sep 2020. ISSN 02765624. doi: 10.1016/j.rssm.2020.100546. URL https://linkinghub.elsevier.com/retrieve/pii/S0276562420300822.

Jing Liu and Susanna Loeb. Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, pages 1216–8430R3, 7 2019. ISSN 0022-166X. doi: 10.3368/jhr.56.2.1216-8430r3. URL http://jhr.uwpress.org/content/early/2019/07/

02/jhr.56.2.1216-8430R3http://jhr.uwpress.org/content/early/2019/07/02/jhr.56.2.1216-8430R3.abstract.

Mary C. Murphy, Maithreyi Gopalan, Evelyn R. Carter, Katherine T.U. Emerson, Bette L. Bottoms, and Gregory M. Walton. A customized belonging intervention improves retention of socially disadvantaged students at a broad-access university. *Science Advances*, 6:eaba4677, 7 2020. ISSN 23752548. doi: 10.1126/sciadv.aba4677. URL http://advances.sciencemag.org/.

Hessel Oosterbeek, Nienke Ruijs, and Inge de Wolf. Using admission lotteries to estimate heterogeneous effects of elite schools, 2020. URL https://www.dropbox.com/s/1p4t9cxhosctyrs/eliteschools2020march.pdf?dl=0.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics*, 37:187–204, 4 2019. ISSN 15372707. doi: 10.1080/07350015.2016.1227711.

Nathan Petek and Nolan G Pope. The multidimensional impact of teachers on students, 2020.

Tim R. Sass, Ron W. Zimmer, Brian P. Gill, and T. Kevin Booker. Charter High Schools' Effects on Long-Term Attainment and Earnings. *Journal of Policy Analysis and Management*, 35(3):683–706, jun 2016.

Ying Shi. Who benefits from selective education? evidence from elite boarding school admissions. *Economics of Education Review*, 74:101907, 2 2020. ISSN 02727757. doi: 10.1016/j.econedurev.2019.07.001.

Victoria F. Sisk, Alexander P. Burgoyne, Jingze Sun, Jennifer L. Butler, and Brooke N. Macnamara. To what extent and under which circumstances are growth mind-sets important to academic achievement? two meta-analyses. *Psychological Science*, 29:549–571, 4 2018. ISSN 0956-7976. doi: 10.1177/0956797617739704. URL http://journals.sagepub.com/doi/10.1177/0956797617739704.

C Spearman. The proof and measurement of association between two things. *Source: The American Journal of Psychology*, 15:72–101, 1904.

C. Spearman. The proof and measurement of association between two things. by c. spearman, 1904. *The American journal of psychology*, 100:441–471, 1987. doi: 10.2307/1422689.

Christopher R. Walters. The demand for effective charter schools. *Journal of Political Economy*, 126:2179–2223, 12 2018. ISSN 1537534X. doi: 10.1086/699980. URL https://www.journals.uchicago.edu/doi/abs/10.1086/699980.

Gregory M. Walton and Geoffrey L. Cohen. A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, 92:82–96, 1 2007. ISSN 00223514. doi: 10.1037/0022-3514.92.1.82.

Gregory M. Walton and Geoffrey L. Cohen. A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, 331:1447–1451, 3 2011. ISSN 00368075. doi: 10.1126/science.1198364. URL https://pubmed.ncbi.nlm.nih.gov/21415354/.

Tables and Figures

Table 1: Summary Statistics

	Analytic Sample		Bottom Decile of Educational Advantage		Top Decile of Educational Advantage	
	mean	SD	mean	SD	mean	SD
			Demographics			
Female	0.4916	0.4999	0.2383	0.4261	0.6997	0.4584
Special education (IEP)	0.1834	0.3870	0.4573	0.4982	0.0555	0.2290
Free lunch	0.7879	0.4088	0.9493	0.2193	0.4462	0.4971
Reduced-price lunch	0.0734	0.2608	0.0188	0.1360	0.1551	0.3620
Census Block SES	-0.4616	0.8658	-0.5865	0.8155	-0.1175	0.9053
White	0.0847	0.2784	0.0398	0.1955	0.2296	0.4206
Black	0.4121	0.4922	0.5247	0.4994	0.2409	0.4276
Native American	0.0017	0.0417	0.0020	0.0451	0.0029	0.0540
Asian/Pacific Islander	0.0325	0.1772	0.0013	0.0365	0.1883	0.3910
Latino	0.4589	0.4983	0.4301	0.4951	0.3309	0.4706
			9th grade Intemediate Outcomes			
Test Scores in 9th Grade	-0.0276	0.9834	-1.0069	0.6406	1.4116	0.7158
Work Hard in 9th Grade	0.1795	0.9874	-0.0821	1.0236	0.5128	0.9693
Social in 9th Grade	-0.0026	0.9988	-0.2380	1.0476	0.3321	0.9793
Surveys in 9th Grade	0.1718	0.9523	-0.1003	0.9952	0.5350	0.9287
Behavior in 9th Grade	0.1688	0.7620	-0.5190	1.4222	0.4531	0.1963
Days Absent in 9th Grade	15.1211	18.7236	34.2316	27.7771	5.5731	7.6340
Days Suspended in 9th grade	0.8183	3.3172	2.9710	6.6585	0.0625	0.6703
Diciplinary Incidents in 9th Grade	0.0782	0.4218	0.2927	0.8692	0.0068	0.0945
			8th Grade Measures			
Math in 8th Grade	0.1908	0.9377	-0.8721	0.6083	1.7914	0.7449
ELA in 8th Grade	0.1959	0.9355	-0.8622	0.8200	1.6068	0.7917
Emotional Health in 8th Grade	0.0673	0.8972	-0.1983	0.8928	0.3199	0.9308
Academic Engagement in 8th Grade	0.2691	0.9137	0.1252	0.8593	0.3703	1.0132
Grit in 8th Grade	0.0440	0.8373	-0.3126	0.8768	0.4289	0.8166
School Connectedness in 8th Grade	0.1393	0.9015	-0.0192	0.8711	0.4281	0.9851
Study Habits in 8th Grade	0.1497	0.8904	-0.2339	0.8531	0.6815	0.9567
Absences in 8th Grade	8.7303	8.6344	19.5467	12.4657	4.5945	3.9139
GPA in 8th Grade	2.7899	0.7795	2.0345	0.7813	3.6004	0.4903
Days Suspended in 8th Grade	0.4479	1.8229	2.2767	4.4812	0.0230	0.2606
Incidents in 8th Grade	0.0655	0.3359	0.3843	0.8494	0.0011	0.0349
Long-term Outcomes						
Any school-Based arrest	0.0377	0.1905	0.1254	0.3312	0.0044	0.0662
Graduation	0.7392	0.4391	0.4300	0.4951	0.9359	0.2449
Enrolled in any college within 2 years	0.5288	0.4992	0.1740	0.3791	0.8711	0.3351
Enrolled in a 4 year college within 2 years	0.3386	0.4732	0.0596	0.2368	0.7758	0.4171
Enrolled in a 2 year college within 2 years	0.2764	0.4472	0.1286	0.3348	0.2499	0.4330
N	157081		15709		15709	

Notes: Number of observations may vary by variable due to missingness and variation in cohorts for which a variable was collected. For more information see Appendix Table 15

Table 2: Temporal Stability of Value-Added and Correlations Across Value-Added

Correlations of Value-Added Within Outcomes Across Time

	Test-Score Value Added	Social Value Added	Work Hard Value Added	Behavior Value Added
t+1	0.4265	0.5475	0.3990	0.5894
t+2	0.1251	0.3891	0.2789	0.5176
t+3	0.3774	0.3358	0.2541	0.4555
t+4	0.4435	0.3478	0.2820	0.2873

Correlations of Average School-Level Value-Added Across Outcomes (143 Schools)

Test Scores Value Added	1.000			
Social Value Added	0.520	1.000		
Work Hard Value Added	0.184	0.346	1.000	
Behavior Value Added	0.353	0.200	0.007	1.000

Disattenuated Correlations of Average School-Level Value-Added Across Outcomes (143 Schools)

Test Scores Value Added	1.000			
Social Value Added	0.820	1.000		
Work Hard Value Added	0.313	0.680	1.000	
Behavior Value Added	0.435	0.285	0.011	1.000

Notes: All reported results are restricted to school-year cells with at least 10 respondents. The **top panel** reports, for each 9^{th} grade measure measure, the correlations between a schools 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-addeds (estimated across all years) for the 9^{th} grade measures.

Table 3: Average Impacts of School Effectiveness and Value-Added

	1	2	3	4	5	6	7	8	9
	Test scores 9th Grade	Surveys 9th Grade	Behaviors 9th Grade	HS Graduation	Enrolled in Any College Within 2 Years	School-Based Arrests	Enrolled in 4-Year College Within 2 Years	Enrolled in 2-Year College Within 2 Years	Persists in College After 1 Year
School Effectiveness Index	0.0886***	0.0927***	0.0629***	0.0247***	0.0261***	-0.00865***	0.0430***	-0.00709	0.0199***
	(0.0125)	(0.00814)	(0.0118)	(0.00400)	(0.00589)	(0.00254)	(0.00854)	(0.00442)	(0.00502)
Socioemotional Value-Added	0.0630***	0.0805***	0.0353***	0.0203***	0.0192***	-0.00622***	0.0331***	-0.00676*	0.0145***
	(0.0124)	(0.00980)	(0.0109)	(0.00389)	(0.00537)	(0.00228)	(0.00850)	(0.00397)	(0.00455)
Test-Score Value-Added	0.0678***	0.0352***	0.0252***	0.00935***	0.0149***	-0.00365**	0.0231***	-0.00215	0.0126***
	(0.0123)	(0.00733)	(0.00711)	(0.00311)	(0.00517)	(0.00146)	(0.00654)	(0.00402)	(0.00461)
Behavior Value-added	0.0370***	0.0486***	0.198***	0.0174***	0.0212***	-0.0144***	0.0315***	-0.00452	0.0164***
	(0.0126)	(0.0133)	(0.0127)	(0.00552)	(0.00541)	(0.00345)	(0.00652)	(0.00407)	(0.00399)

Robust standard errors in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Notes: Each point estimate comes from a separate regression

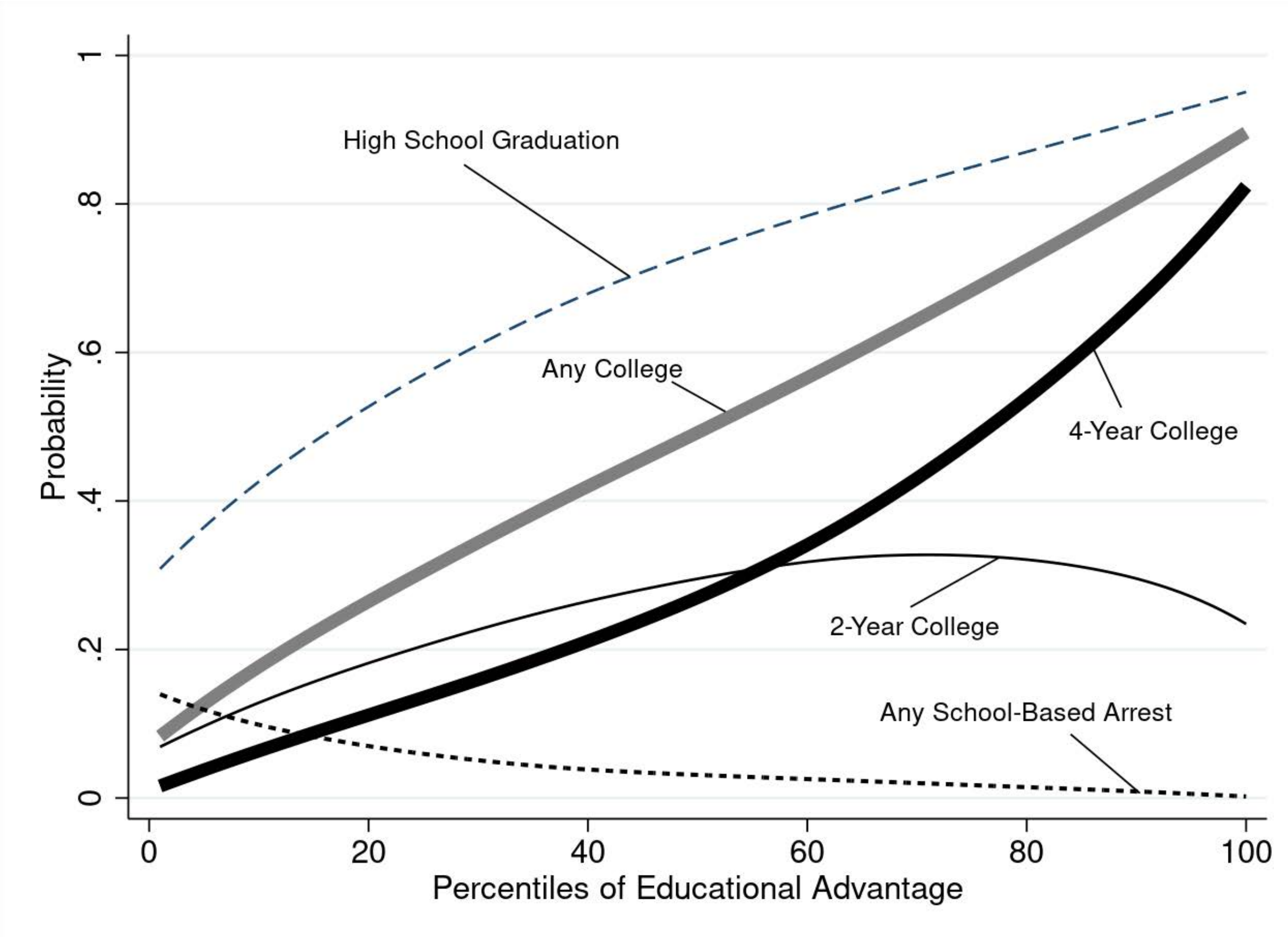
Results are based on regression of outcomes on a single measure of out-of-sample school impacts (overall effectiveness, test score value-added, socio-emotional value-added, or behaviour value-added). All models include individual demographic controls (race / ethnicity, free and reduced price lunch, and gender), 8th grade lags (math and ELA test scores, survey measures, absences, and discipline), and school-level averages for all the demographics and lagged measures, as well as year fixed effects. We also include the socio-economic status of the student census block proxied by average occupation status and education levels. Missing 8^{th} grade measures were imputed using 7^{th} grade measures and demographic characteristics. For the longer-run college outcomes, the sample includes first time 9^{th} grade students between 2011 and 2014. For the longer-run high-school outcomes, the sample includes first time 9^{th} grade students between 2011 and 2015. For the measures, the sample includes first time 9^{th} grade students between 2011 and 2017. **Note:** Sample sizes may differ across outcomes due to some missingness in 9th grade test scores and surveys.

Table 4: Testing for Selection

	Intermediate Outcomes				Main Longer-Run Outcomes				Detailed College Outcomes			
	1	2	3	4	5	6	7	8	9	10	11	12
	9th Grade Test Scores			Predicted	HS Graduation			Predicted	Enrolled in 4 Year College within 2 Years			Predicted
School Effectiveness Index	0.0886***	0.0447***	0.0580***	0.000268	0.0247***	0.0388***	0.0172**	0.00147	0.0430***	0.0397***		0.000158
	(0.0125)	(0.0172)	(0.0127)	(0.00335)	(0.00400)	(0.0114)	(0.00842)	(0.00114)	(0.00854)	(0.0139)		(0.00396)
Observations	102,235	99,683	16,386	160,148	82,146	79,550	8,188	160,148	55,564	53,242	3,399	160,148
F-statistic on First Stage		310.9				396.8				364.4		
Number of Families			7786				3902				1634	
	9th Grade Survey Measures			Predicted	In-school Arrests			Predicted	Enrolled in 2 Year College within 2 Years			Predicted
School Effectiveness Index	0.0927***	0.105***	0.0573***	0.00557	-0.00865***	-0.0184***	-0.00948**	-0.000627	-0.00709	-0.0214*		0.000483
	(0.00814)	(0.0166)	(0.0155)	(0.00701)	(0.00254)	(0.00583)	(0.00477)	(0.000856)	(0.00442)	(0.0121)		(0.00183)
Observations	124,867	122,104	28,804	160,148	82,146	79,550	8,188	160,148	55,564	53,242	3,399	160,148
F-statistic on First Stage		281.5				396.8				364.4		
Number of Families			13584				3902				1634	
	9th Grade Behaviors			Predicted	Enrolled in Any College within 2 Years			Predicted	Persist In College After 1 Year			
School Effectiveness Index	0.0629***	0.0841***	0.0341***	0.00702	0.0261***	0.00745		0.000728	0.0199***	0.0116		0.000904
	(0.0118)	(0.0192)	(0.00987)	(0.00755)	(0.00589)	(0.0152)		(0.00172)	(0.00502)	(0.0138)		(0.00152)
Observations	157,628	153,966	41,711	160,148	55,560	53,239	3,399	160,148	55,564	53,242	3,399	160,148
F-statistic on First Stage		310.7				364.4				364.4		
Number of Families			19420				1634				1634	
Sibling FE			X				X				X	
School Assignment IV		X				X				X		

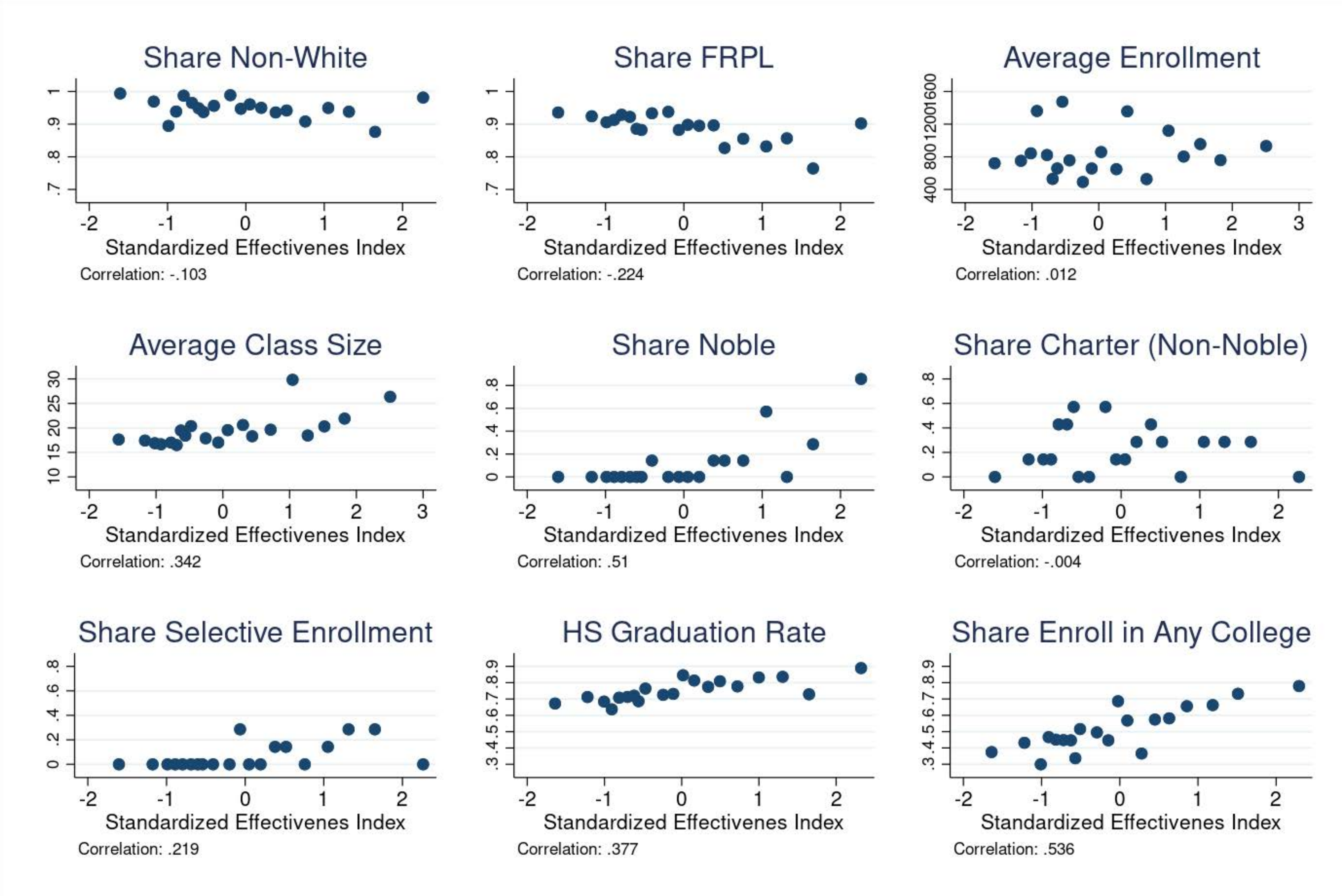
Robust standard errors adjusted for clustering at the school level.
*** p<0.01, ** p<0.05, * p<0.1
Results are based on regression of outcomes on out-of-sample school effectiveness. All models include individual demographic controls (race / ethnicity, free and reduced price lunch, and gender), 8th grade lags (math and ELA test scores, survey measures, absences, and discipline), and school-level averages for all the demographics and lagged measures, as well as year fixed effects. We also include the socio-economic status of the student census block proxied by average occupation status and education levels. Missing 8^{th} grade measures were imputed using 7^{th} grade measures and demographic characteristics. For the longer-run outcomes, the sample includes first time 9^{th} grade students between 2011 and 2014. For the measures, the sample includes first time 9^{th} grade students between 2011 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., 8^{th} grade test scores, surveys, behaviours), gender, ethnicity, free-lunch status, and the socio-economic status 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-addeds with only school-average math scores, school average absences and the percent white at the school. **Note:** Sample sizes may differ across outcomes due to some missingness in 9th grade test scores and surveys.

Figure 1. *Average Outcomes: By Estimated Educational Advantage*



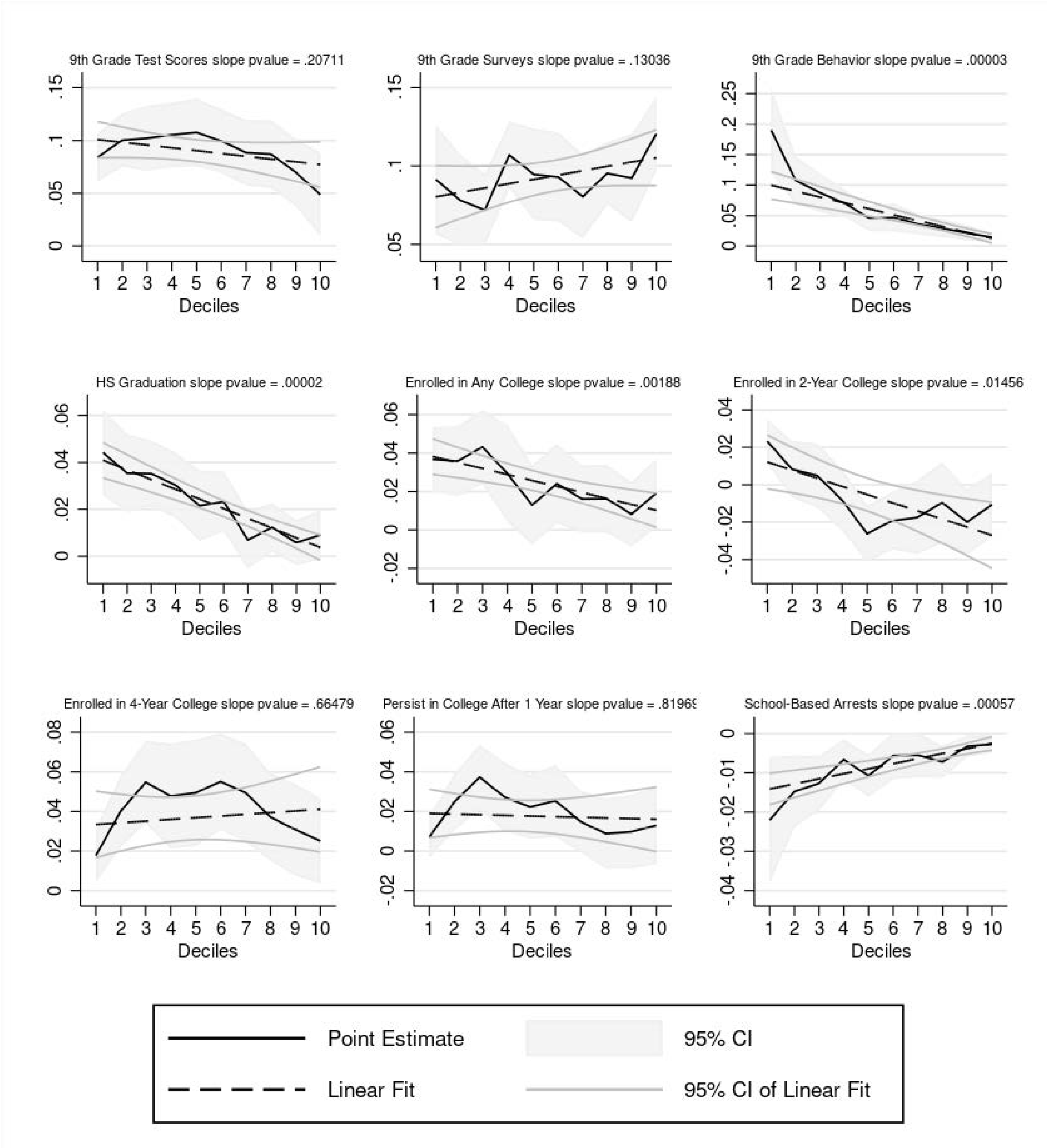
Notes: This figures plots the average of each outcome for different percentiles of the estimated educational advantage distribution. The predicted educational advantage is the fitted value from an ordered probit model predicting the level of education attained based on all 8^{th} grade measures and demographics (*in all other years*). We present the coefficient estimates from the ordered probit model for the full sample in Appendix Table A3. We also present plots of outcome educational advantage within race and gender groups in Appendix Figure A1.

Figure 2. *Difference in Characteristics Among Most and Least Effective Schools*



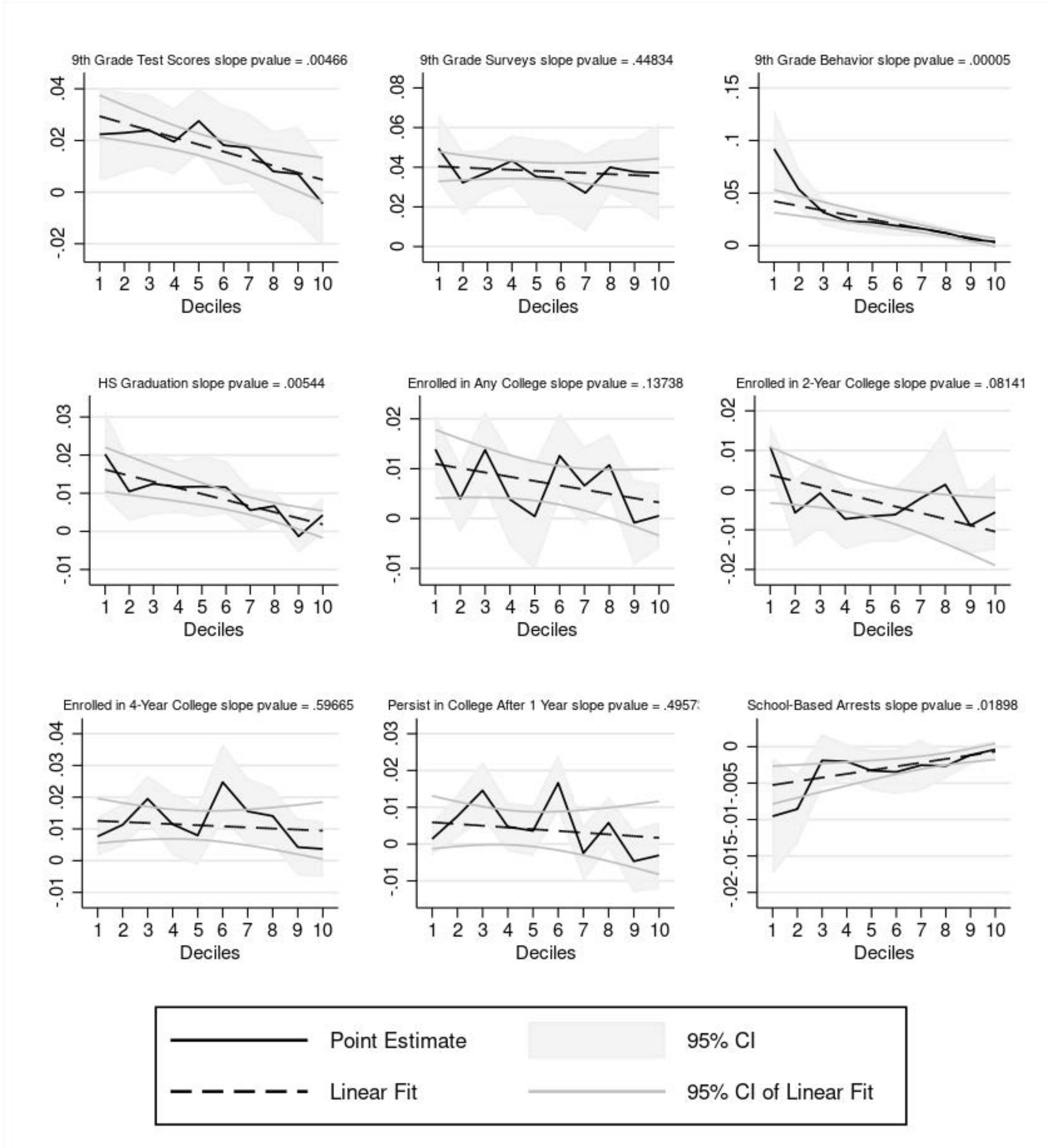
Notes: Each panel is a binned scatter-plot (20 bins) of the school characteristic against the standardized estimated school effectiveness. We report the raw correlation between the school characteristic and effectiveness below each plot for a sample of 131 schools. Share FRPL is defined as the share 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.

Figure 3. *Impacts on Outcomes: By Estimated Educational Advantage*



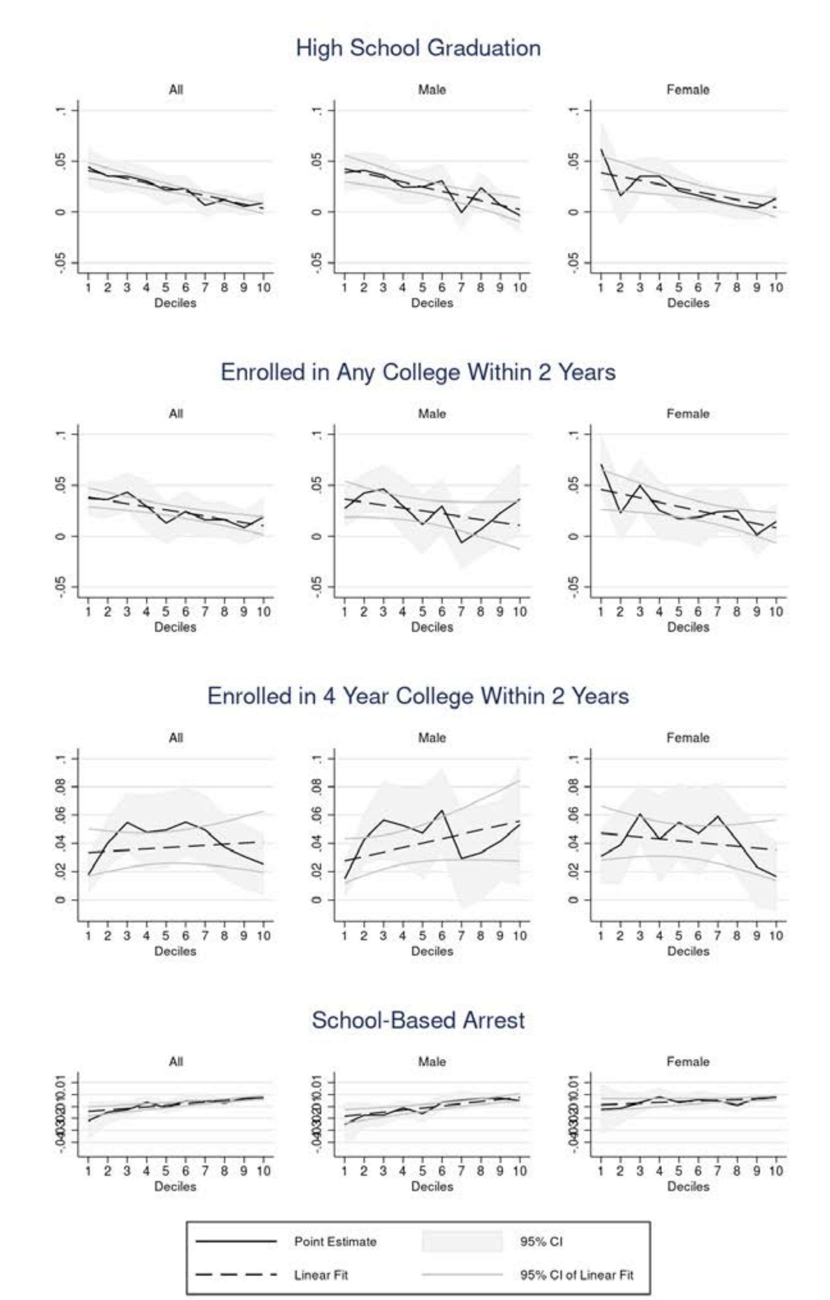
Notes: Each graph represents the marginal impact of a 1 standard deviation increase in overall school effectiveness for different deciles of the educational advantage distribution for a single outcome. Each panel presents the results of 10 separate regressions each defined as in Equation (6). The 95 percent confidence interval for each point estimate is depicted by the grey shaded 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 confidence interval.

Figure 4. *Impacts on Outcomes Unexplained by Test Score: By Educational Advantage*



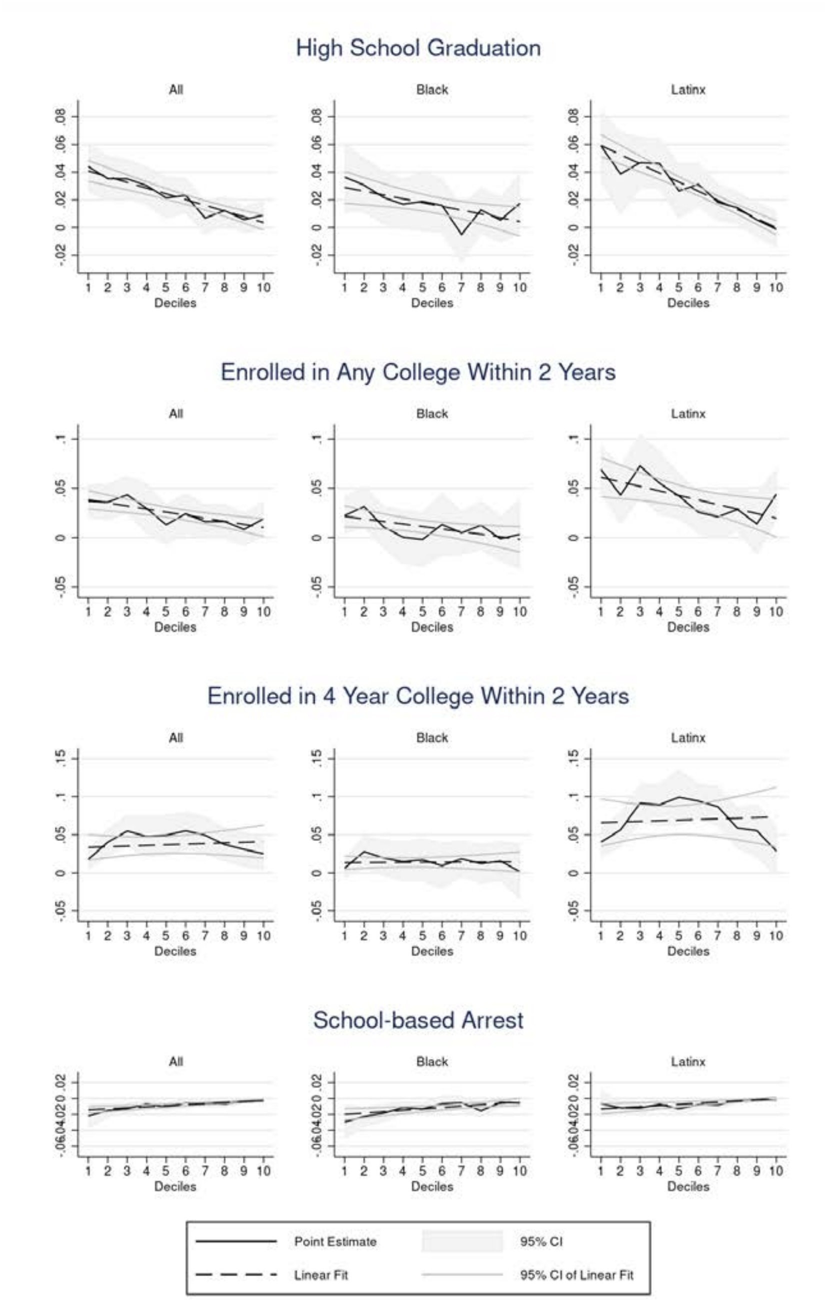
Notes: Each graph depicts the *difference* between the marginal impact of a 1 standard deviation increase in the school effectiveness index and in test-score value-added. That is, it shows the impact of school effectiveness on each outcome that cannot be explained by test scores. Each regression model controls for the same covariates as in Equation (6). Each panel presents the results of 10 separate regressions. The 95 percent confidence interval for each point estimate is depicted by the grey shaded 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 confidence interval.

Figure 5. *Impacts on Outcomes: By Estimated Educational Advantage and Sex*



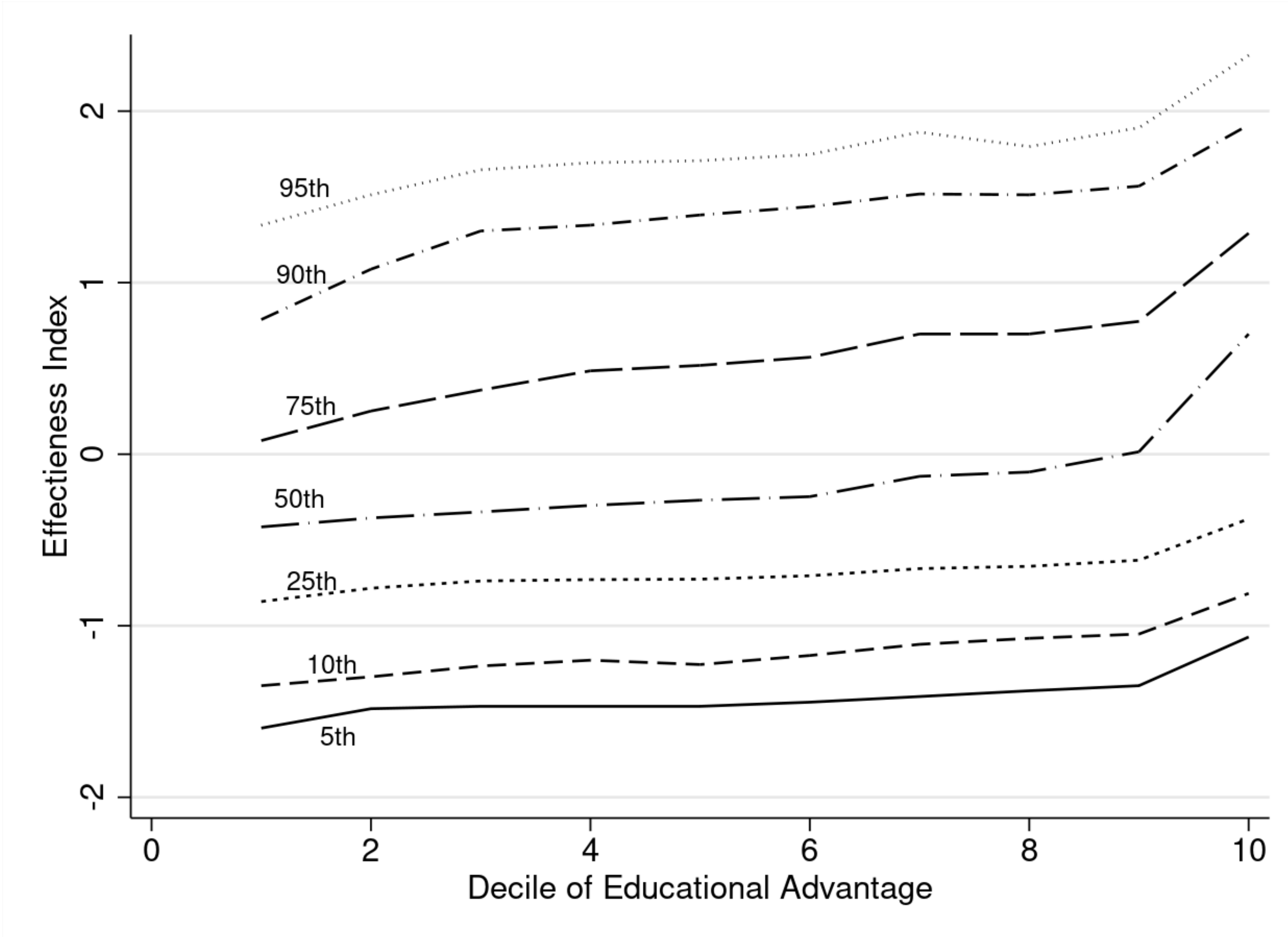
Notes: Each graph represents the marginal impact of a 1 standard deviation increase in overall school effectiveness for different deciles of the educational advantage distribution for a single outcome and sub-population. Each panel presents the results of 10 separate regressions each defined as in Equation (6). The 95 percent confidence interval for each point estimate is depicted by the grey shaded 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 confidence interval.

Figure 6. *Outcomes by Advantage: By Race / Ethnicity*



Notes: 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 and sub-population. Each panel presents the results of 10 separate regressions each defined as in Equation (6). The 95 percent confidence interval for each point estimate is depicted by the grey shaded 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 confidence interval.

Figure 7. *Percentiles of Effectiveness Index: By Estimated Educational Advantage*



Notes: This figure plots the distribution of the overall effectiveness for students with different levels of educational advantage. The lines depict the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentile of school effectiveness that students in each decile are exposed to.

VII Appendix

Table A1: Summary Statistics for Survey Completers and Non-Completers

	Analytic Sample		Completed the Surveys		Did not complete Surveys	
	mean	SD	mean	SD	mean	SD
Demographics						
Female	0.4916	0.4999	0.502669	0.499995	0.458725	0.4983
Special education (IEP)	0.1834	0.3870	0.158572	0.365278	0.258087	0.437588
Free lunch	0.7879	0.4088	0.780976	0.413587	0.808699	0.393331
Reduced-price lunch	0.0734	0.2608	0.077028	0.266637	0.0625	0.242065
Census Block SES	-0.4616	0.8658	-0.46797	-0.87357	-0.44255	-0.84185
White	0.0847	0.2784	0.089733	0.2858	0.069668	0.254591
Black	0.4121	0.4922	0.382196	0.485926	0.50176	0.500003
Native American	0.0017	0.0417	0.001672	0.040855	0.001939	0.043989
Asian/Pacific Islander	0.0325	0.1772	0.036384	0.187244	0.020714	0.142428
Latino	0.4589	0.4983	0.480416	0.499618	0.394413	0.488731
9th grade Intemediate Outcomes						
Test Scores in 9th Grade	-0.0276	0.9834	0.029924	-0.96869	-0.21285	-1.00746
Work Hard in 9th Grade	0.1795	0.9874	0.186834	-0.98092	-0.02135	-1.13492
Social in 9th Grade	-0.0026	0.9988	0.003019	-0.99426	-0.15024	-1.09945
Surveys in 9th Grade	0.1718	0.9523	0.179159	-0.94566	-0.01489	-1.08968
Behavior in 9th Grade	0.1688	0.7620	0.233114	-0.64323	-0.029	-1.02141
Days Absent in 9th Grade	15.1211	18.7236	12.99633	15.59982	21.65436	24.95769
Days Suspended in 9th grade	0.8183	3.3172	0.644835	2.795615	1.337551	4.495297
Diciplinary Incidents in 9th Grade	0.0782	0.4218	0.061845	0.3596	0.127143	0.566646
On Track in 9th Grade	0.8462	0.3607	0.870445	0.335815	0.757014	0.428896
8th Grade Measures						
Math in 8th Grade	0.1908	0.9377	0.25101	-0.93307	0.010372	-0.92871
ELA in 8th Grade	0.1959	0.9355	0.257527	-0.91814	0.010957	-0.96243
Emotional Health in 8th Grade	0.0673	0.8972	0.079809	-0.90438	0.029781	-0.87456
Academic Engagement in 8th Grade	0.2691	0.9137	0.275683	-0.92486	0.249519	-0.87944
Grit in 8th Grade	0.0440	0.8373	0.052673	-0.84616	0.017878	-0.81006
School Connectedness in 8th Grade	0.1393	0.9015	0.143819	-0.91049	0.125375	-0.87399
Study Habits in 8th Grade	0.1497	0.8904	0.16246	-0.90448	0.111173	-0.84576
Absences in 8th Grade	8.7303	8.6344	8.113539	7.758448	10.57956	10.63503
GPA in 8th Grade	2.7899	0.7795	2.837592	0.772915	2.646929	0.781757
Days Suspended in 8th Grade	0.4479	1.8229	0.360557	1.558484	0.709553	-2.43069
Incidents in 8th Grade	0.0655	0.3359	0.053284	0.287962	0.102219	0.448415
Long-Run Outcomes						
Any school-Based arrest	0.0377	0.1905	0.031782	0.175422	0.053904	0.225834
Graduation	0.7392	0.4391	0.777252	0.416094	0.63654	0.481007
Enrolled in any college within 2 years	0.5288	0.4992	0.573986	0.494502	0.405241	0.490956
Enrolled in a 4 year college within 2 years	0.3386	0.4732	0.373327	0.483694	0.243449	0.429179
Enrolled in a 2 year college within 2 years	0.2764	0.4472	0.29627	0.456617	0.222361	0.415846
N	157027		117827		39200	

Notes: Survey completers are students who have 9^{th}-grade data for emotional health, academic engagement, grit, school connectedness, and study habits. As such, we report averages for some measures even among non-completers because many noncompleters are missing some data but not others.

Table A2: Psychometric Properties of SED measures (*as reported by the University of Chicago Consortium on School Research*): 2011 through 2013

Measure	**School Year**	**Separation**	**Reliability**	**Item Infits**	**Item Outfits**
Grit	2010-11	1.68	0.74	0.84, 0.76, 0.71, 1.24	0.85, 0.76, 0.71, 1.19
Social Skills	2010-11	1.69	0.74	1.08, 1.36, 1.41, 1.11	1.05, 1.33, 1.44, 1.15
Academic Effort	2010-11	1.74	0.75	0.85, 1.22, 1.1, 0.91	0.82, 1.17, 1.12, 0.94
Academic Engagement	2010-11	1.59	0.7	0.49, 0.56, 0.71, 0.56	0.49, 0.57, 0.72, 0.58
Belonging	2010-11	2.07	0.81	0.93, 1.02, 0.99, 0.96, 1.29	0.91, 0.97, 0.99, 0.93, 1.33
Grit	2011-12	1.54	0.7	0.8, 0.73, 0.68, 1.19	0.81, 0.57, 0.6, 0.42
Social Skills	2011-12	1.68	0.74	1.37, 1.36, 1.28, 1.06	1.68, 1.24, 1.18, 0.95
Academic Effort	2011-12	1.75	0.75	0.85, 1.22, 1.08, 0.92	0.82, 1.17, 1.1, 0.96
Academic Engagement	2011-12	1.56	0.71	0.54, 0.53, 0.47, 0.69	0.56, 0.55, 0.48, 0.71
Belonging	2011-12	2.13	0.82	0.98, 1.28, 0.91, 1.02, 0.97	0.97, 1.32, 0.89, 0.97, 0.94
Grit	2012-13	1.55	0.71	0.77, 0.69, 0.63, 1.13	0.79, 0.7, 0.63, 1.1
Social Skills	2012-13	1.67	0.74	1.3, 1.37, 1.23, 1.04	1.55, 1.25, 1.12, 0.94
Academic Effort	2012-13	1.77	0.76	0.86, 1.2, 1.13, 0.94	0.83, 1.15, 1.15, 0.97
Academic Engagement	2012-13	1.57	0.71	0.55, 0.54, 0.47, 0.69	0.57, 0.56, 0.48, 0.70
Belonging	2012-13	2.14	0.82	0.95, 1.28, 0.90, 1.03, 0.96	0.95, 1.31, 0.87, 0.98, 0.93

Notes. The reported statistics are from internal documentation at the University of Chicago Consortium on School Research where Rasch analysis was performed on individual survey items. All measures are anchored to 2010-11 step and item difficulties. Infit and outfit measures greater than 1 indicate underfit to the Rasch model and values lower than 1 indicate overfit. Generally, infit and outfit values in the range of 0.6-1.4 are considered reasonable for survey measures. Reliability represents individual reliability and includes extreme people. The patterns are very similar for years 2013 through 2018.

Table A3: Ordered Probit Parameter Estimates

	Educational Advantage		cont'd
8th Grade Math	0.301***	Native	-0.526**
	(0.0105)		(0.212)
8th Grade Math Squared	0.00557	Asian	-0.0345
	(0.00588)		(0.165)
8th Grade ELA	0.172***	Latinx	-0.389**
	(0.00934)		(0.163)
8th Grade ELA Squared	0.0154***	Other Race	-0.0335
	(0.00476)		(0.349)
Emotional Health in 8th Grade	-0.0118	Female	0.0383
	(0.00720)		(0.151)
Academic Engagement in 8th Grade	-0.000894	Female * White	0.111
	(0.00619)		(0.159)
Grit in 8th Grade	0.0420***	Female * Black	0.313**
	(0.00452)		(0.156)
School Connectedness in 8th Grade	-0.0250***	Female * Native	0.351*
	(0.00727)		(0.188)
Study Habits in 8th Grade	0.106***	Female * Asian	0.153
	(0.00840)		(0.156)
8th Grade Top 25% of Absences	-0.574***	Female * Latinx	0.198
	(0.0153)		(0.150)
Serious Incidents in 7th or 8th Grade	-0.392***	Female * Other Race	-0.531
	(0.0285)		(0.422)
Receive Free Lunch	-0.161***	/cut1	-1.204***
	(0.0615)		(0.184)
Receive Reduced Price Lunch	0.0754	/cut2	-0.518***
	(0.0610)		(0.190)
White	-0.309**	/cut3	0.0684
	(0.147)		(0.196)
Black	-0.404**		
	(0.178)	Observations	116,162

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Note that the sample size is larger than the analytic sample used for the main outcome analysis. This is prediction model uses all available data, which include observation for individuals who attend schools that do not have valid value-added estimates. the results are very similar is we restrict the prediction to only those same individual in the main analytic long term sample.

Table A4: Effect of SED Value-Added on Average Intermediate and Long-Term Student Outcomes

	(1) Test scores 9th Grade	(2) Surveys 9th Grade	(3) Behaviors 9th Grade	(4) HS Graduation	(5) School-Based Arrests	(6) Enrolled in Any College Within 2 Years	(7) Enrolled in 4-Year College Within 2 Years	(8) Enrolled in 2-Year College Within 2 Years	(9) Persists in College After 1 Year
Workhard Value-Added	0.0601***	0.0732***	0.0297***	0.0203***	-0.00641***	0.0204***	0.0341***	-0.00639	0.0161***
	(0.0135)	(0.0101)	(0.0110)	(0.00442)	(0.00232)	(0.00575)	(0.00942)	(0.00419)	(0.00488)
Social Value-Added	0.0735***	0.0955***	0.0468***	0.0220***	-0.00591**	0.0196***	0.0345***	-0.00715	0.0136***
	(0.0124)	(0.00995)	(0.0131)	(0.00441)	(0.00295)	(0.00575)	(0.00847)	(0.00446)	(0.00494)
Observations	102,235	124,867	157,628	82,146	82,146	55,560	55,564	55,564	55,564

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Results are based on separate regressions of outcomes on out-of-sample socioemotional value-added, disaggregated into the social well-being and work hard constructs. All models include individual demographic controls (race / ethnicity, free and reduced price lunch, and gender), 8^{th} grade lags (math and ELA test scores, survey measures, absences, and discipline), and school-level averages for all the demographics and lagged measures, as well as year fixed effects. We also include the socio-economic status of the student census block proxied by average occupation status and education levels. Missing 8^{th} grade measures were imputed using 7^{th} grade measures and demographic characteristics. For the longer-run college outcomes, the sample includes first time 9^{th} grade students between 2011 and 2014. For the longer-run high-school outcomes, the sample includes first time 9^{th} grade students between 2011 and 2015. For the measures, the sample includes first time 9^{th} grade students between 2011 and 2017. Sample sizes may differ across outcomes due to some missingness in 9^{th} grade test scores and surveys. Each point estimate is based on a separate regression.

Table A5: Factor Analysis

	Variance	Difference	Proportion
Factor 1	1.41463	1.38108	1.1886
Factor 2	0.03355	.	0.0282

Rotated factor loadings (pattern matrix) and unique variances

	Factor1	Factor2	Uniqueness
Work Hard Value-Added	0.7922	-0.0144	0.3723
Social Value-Added	0.7954	0.0142	0.3671
Test Scores Value-Added	0.3351	-0.1141	0.8747
Behaviours Value-Added	0.2052	0.1419	0.9378

Method: principal factors

Rotation: orthogonal varimax (Kaiser off).

Note: The proportion explained by this factor is greater than one because the model also includes factors with negative eigenvalues.

Table A6: Results Using Small Within-Family College Sample

	(1) Test Scores	(2) Surveys	(3) Behaviors	(4) Dropout	(5) Graduate
School Effectiveness Index	-0.0178	-1.65e-05	0.0383	0.00320	-0.0110
	(0.0235)	(0.0261)	(0.0284)	(0.00645)	(0.0112)
Observations	1,943	2,439	3,357	3,399	3,399
Sibling FE	X	X	X	X	X
Number of Families	940	1178	1614	1634	1634

	(6) Ever Arrested	(7) Enroll in College	(8) Enroll in 4-Yr College	(9) Enroll in 2-Yr College	(10) Persist in College 1-Yr
School Effectiveness Index	-0.00772	0.00675	0.000218	0.00977	0.00716
	(0.00785)	(0.0131)	(0.00996)	(0.0109)	(0.00851)
Observations	3,399	3,399	3,399	3,399	3,399
Sibling FE	X	X	X	X	X
Number of Families	1634	1634	1634	1634	1634

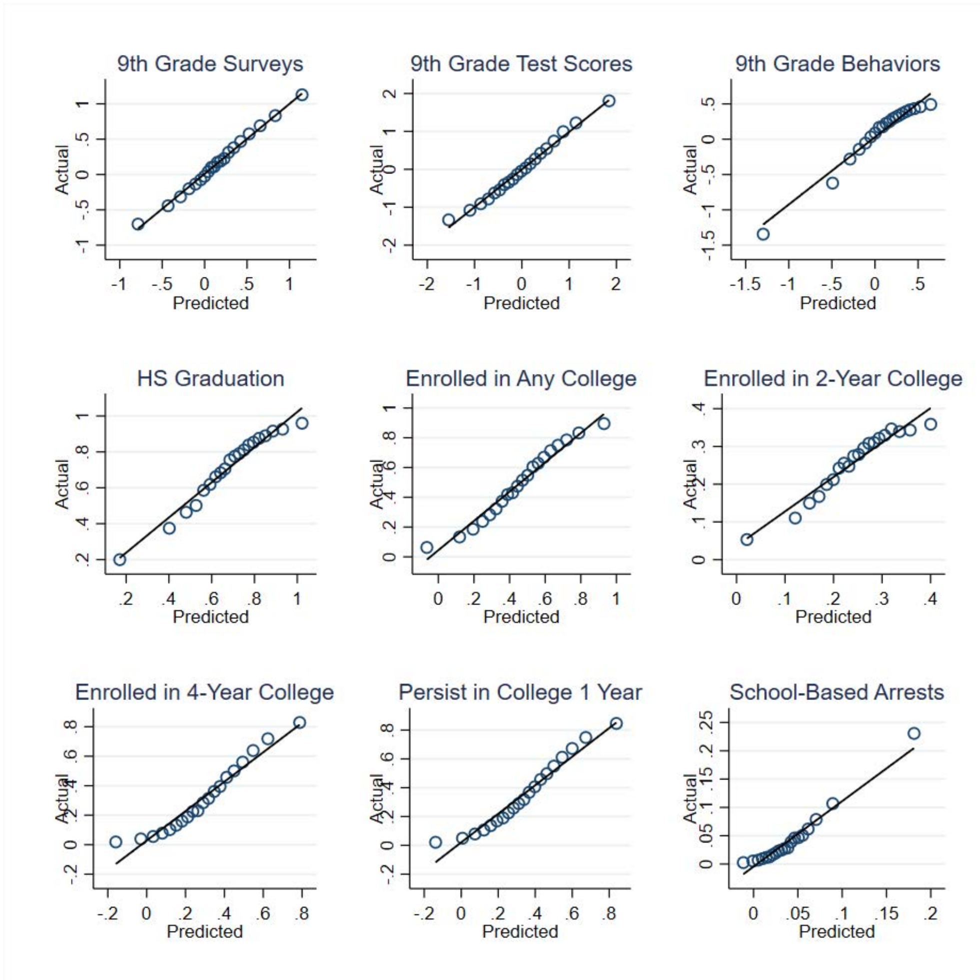
Robust standard errors in parentheses are adjusted for clustering at the school level.
*** p<0.01, ** p<0.05, * p<0.1
Sample: These models are estimated only on the sample of individuals with college outcome data who are linked to other family members who also have college outcome data in our analytic sample. **Model:** 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), 8th grade lags (math and ELA test scores, survey measures, absences, and discipline), and school-level averages for all the demographics and lagged measures, as well as year fixed effects. We also include the socio-economic status of the student census block proxied by average occupation status and education levels. Missing 8^{th} grade measures were imputed using 7^{th} grade measures and demographic characteristics. **These models also include family fixed effects.** For the longer-run outcomes, the sample includes first time 9^{th} grade students between 2011 and 2014. For the measures, the sample includes first time 9^{th} grade students between 2011 and 2017. **Note:** Sample sizes may differ across outcomes due to some missingness in 9th-grade test scores and surveys.

Table A7: Marginal Effects for Select Deciles of Educational Advantage

	Test Scores				Surveys			
	Bottom Decile	Top Decile	Bottom 3 Deciles	Top 3 Deciles	Bottom Decile	Top Decile	Bottom 3 Deciles	Top 3 Deciles
School Effectiveness Index	0.0840***	0.0489**	0.0983***	0.0642***	0.0915***	0.121***	0.0801***	0.101***
	(0.0115)	(0.0201)	(0.0117)	(0.0156)	(0.0177)	(0.0120)	(0.0102)	(0.00914)
	Behaviors				**Grad**			
	Bottom Decile	Top Decile	Bottom 3 Deciles	Top 3 Deciles	Bottom Decile	Top Decile	Bottom 3 Deciles	Top 3 Deciles
School Effectiveness Index	0.190***	0.0145***	0.127***	0.0232***	0.0442***	0.00904*	0.0402***	0.00994***
	(0.0356)	(0.00336)	(0.0225)	(0.00537)	(0.00927)	(0.00537)	(0.00670)	(0.00284)
	College				**Arrested**			
	Bottom Decile	Top Decile	Bottom 3 Deciles	Top 3 Deciles	Bottom Decile	Top Decile	Bottom 3 Deciles	Top 3 Deciles
School Effectiveness Index	0.0368***	0.0192**	0.0409***	0.0147***	-0.0220***	-0.00269**	-0.0161***	-0.00457***
	(0.00853)	(0.00909)	(0.00807)	(0.00535)	(0.00818)	(0.00116)	(0.00483)	(0.00118)
	4-Year College				**2-Year College**			
	Bottom Decile	Top Decile	Bottom 3 Deciles	Top 3 Deciles	Bottom Decile	Top Decile	Bottom 3 Deciles	Top 3 Deciles
School Effectiveness Index	0.0177**	0.0251**	0.0393***	0.0312***	0.0233***	-0.0107	0.0135**	-0.0135**
	(0.00681)	(0.0109)	(0.00821)	(0.00803)	(0.00600)	(0.00876)	(0.00596)	(0.00649)

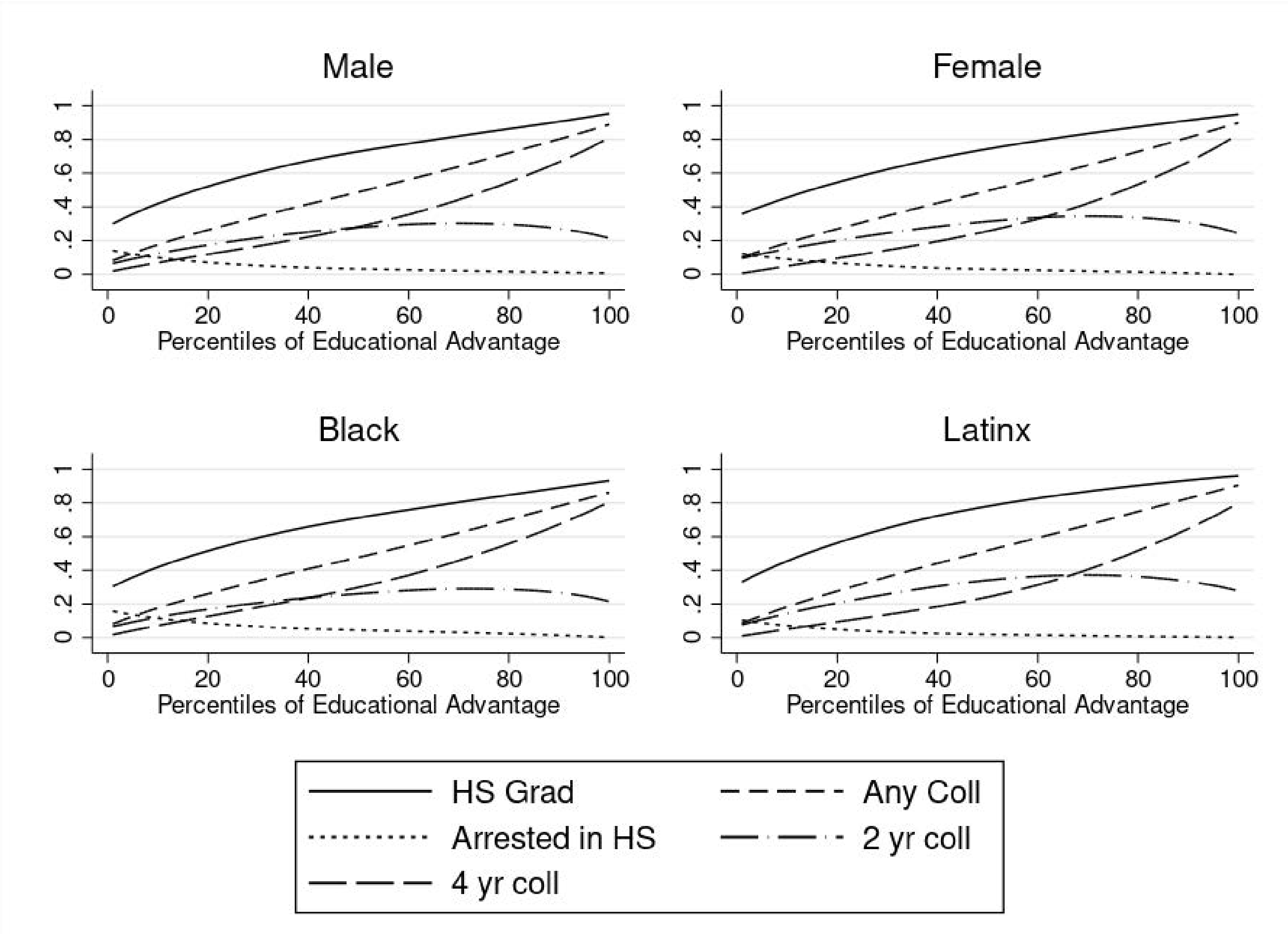
Robust standard errors in parentheses adjusted for clustering at the school level.
*** $p<0.01$, ** $p<0.05$, * $p<0.1$
We report the main result for regression models estimated on different sub-samples of the data. Results are based on regression of outcomes on the out-of-sample overall school effectiveness index. All models include individual demographic controls (race / ethnicity, free and reduced price lunch, and gender), 8th grade lags (math and ELA test scores, survey measures, absences, and discipline), and school-level averages for all the demographics and lagged measures, as well as year fixed effects. We also include the socio-economic status of the student census block proxied by average occupation status and education levels. Missing 8^{th} grade measures were imputed using 7^{th} grade measures and demographic characteristics. For the longer-run college outcomes, the sample includes first time 9^{th} grade students between 2011 and 2014. For the longer-run high-school outcomes, the sample includes first time 9^{th} grade students between 2011 and 2015. For the measures, the sample includes first time 9^{th} grade students between 2011 and 2017.

Figure A1. *Actual Outcome by Predicted Outcome*



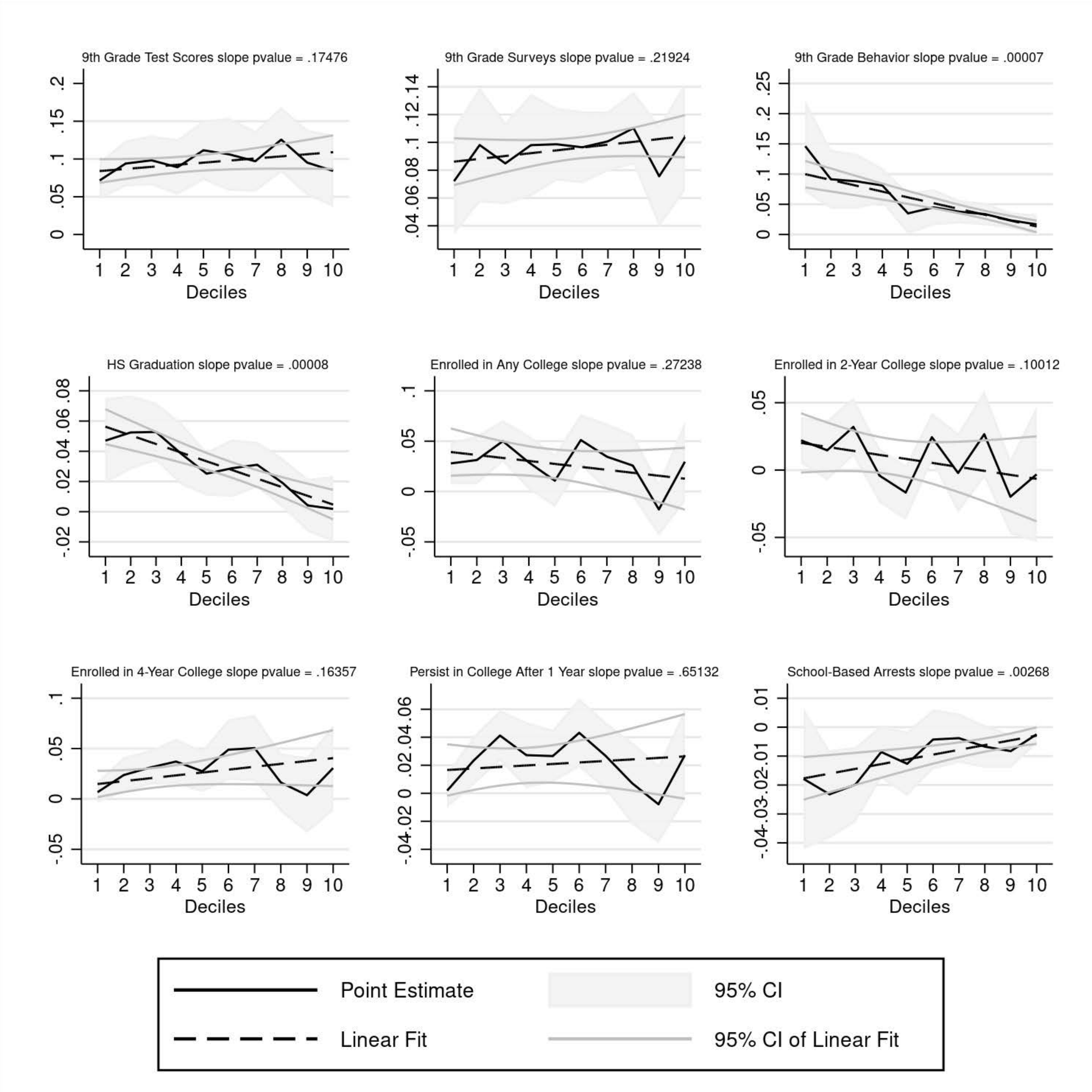
Notes: Each graph presents the average of the actual outcome for different groups of students by predicted outcome. The predicted outcomes are the fitted values from a regression of each outcome on all observed demographics and 8^{th} grade measures based on students in *other* years. The predictors include lagged measures (i.e., 8^{th} grade test scores, surveys, behaviours), gender, ethnicity, free-lunch status, and the socio-economic status of the student's census block.

Figure A2. *Average Outcomes by Educational Advantage: By Race and Gender*



Notes: This figures plots the average of each outcome for different percentiles of the estimated educational advantage distribution by race and gender. The predicted educational advantage is the fitted value from an ordered probit model predicting the level of education attained based on all 8^{th} grade measures and demographics (*in all other years*). We present the coefficient estimates from the ordered probit model for the full sample in Appendix Table A3.

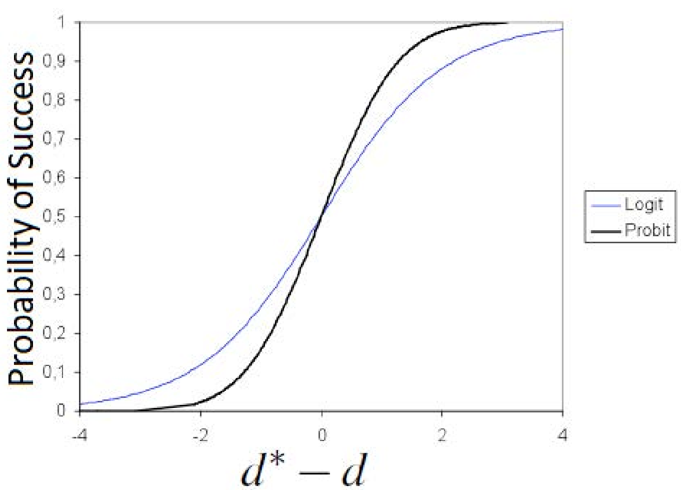
Figure A3. *Impacts on Outcomes: By Educational Advantage (neighborhood schools only)*



Notes: 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) but only on the sample of traditional public school students. The 95 percent confidence interval for each point estimate is depicted by the grey shaded 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 confidence interval.

Appendix B: The Test For Mechanical Heterogeneity

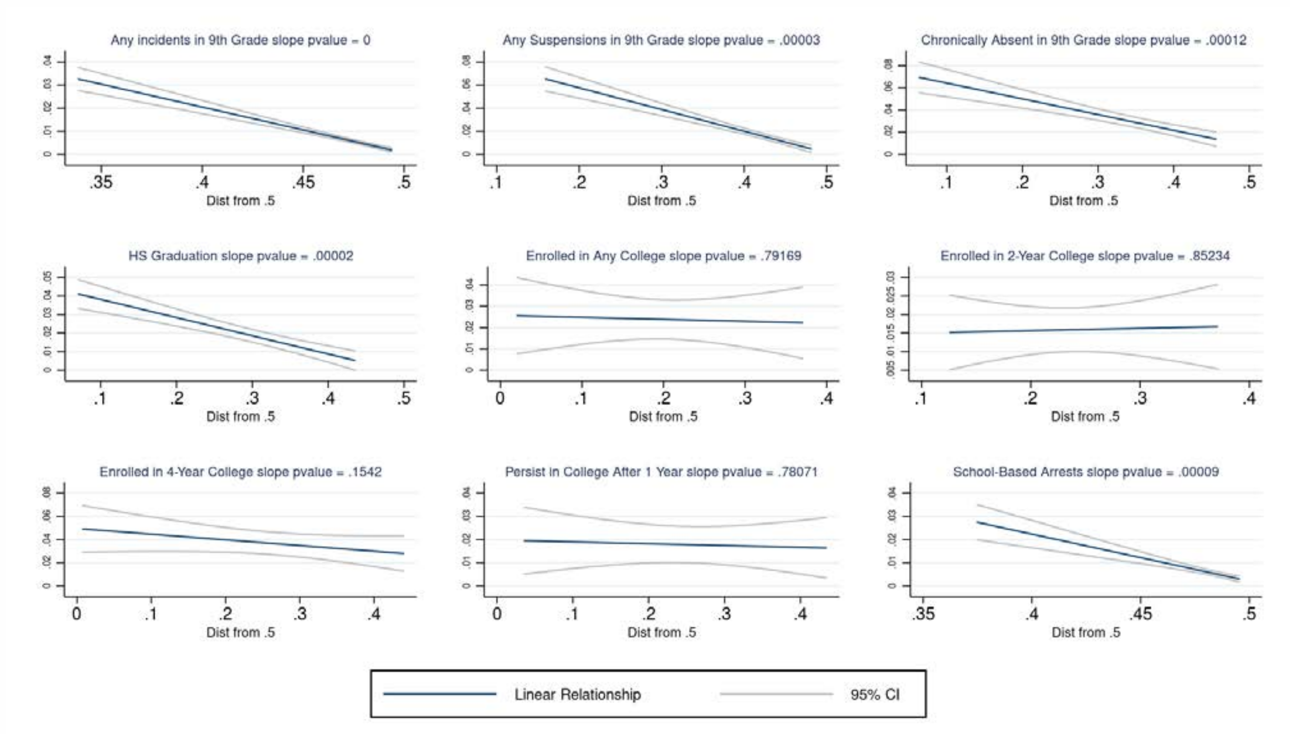
In models such as the logit or probit models, there is some continuous latent variable (d) that summarizes a predisposition to "success" (i.e., a positive outcome). The realized outcome ($D \in 0,1$) is a function of this latent disposition plus some random error, ε, such that $D = 1$ if $d + \varepsilon > d^*$ and $D = 0$ otherwise, where d^* is some unobserved fixed threshold. The probability of success for an individual with disposition d is therefore $Pr(\varepsilon > d^* - d)$. This is a cumulative probability. See the logit and probit models depicted below. The change in the probability due to a marginal increase in d is therefore the probability density of ε at $d^* - d$. Under a symmetric single-peaked bell-shaped distribution of ε (as in a logit or probit model), for the same change in d, the observed change in probability will be largest (in magnitude) for individuals with d close to d^* and success probability close to 0.5, declining in magnitude for those for those with d farther from d^* and success probability farther from 0.5, and smallest for those with d very far from d^* and success probability farthest from 0.5 (i.e., close to 1 or 0).



For example, consider an outcome such as being suspended. The underlying disposition toward success (in this case being suspended) for any group is denoted d_g. Those in the top decile of educational advantage have baseline suspension rates below 1 percent and therefore have large $|d^* - d_g|$ compared to those in the bottom decile who have baseline suspension rates around 30 percent and therefore much smaller $|d^* - d_g|$. The above framework indicates that **with the same change in d_g,** the change in suspension rates will be larger for the bottom decile than the top.

This logic forms the basis for our test. We propose that if the differences in the marginal effect on binary outcomes across groups can be explained by differences in the baseline probabilities across groups, it would be indicative "mechanical heterogeneity". In contrast, if differences in baseline probabilities across groups do not explain differences in marginal effects on these binary outcomes, it would imply that any observed heterogeneity is not mechanical and therefore reflects heterogeneous effects on skills and latent predispositions.

Figure B1. *Relationship Between Probability of Success and Marginal Effects*



Notes: Each graphs plots the linear relationship between the the absolute value of the marginal effect for each decile group and the difference between the average success rate for that same decile group and 0.5. This comes from the regression model laid out and detailed in equation (8). That is, for each binary outcome, we run the regression below

$$|\delta_g| = \alpha + \pi \times (|p_g - 0.5|) + \upsilon_g \tag{9}$$

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. The slope of the plotted lined in each graph is π, which represents the relationship between the absolute value of the marginal effect and the distance between the baseline success rate and 0.5. For each outcome, we report the p-value on the hypotheses that $\pi = 0$. If $\pi = 0$, it would imply that the differences in the marginal effect on binary outcomes can be explained by differences in the baseline probabilities across groups – which would be indicative "*mechanical heterogeneity*". In contrast, if differences in baseline probabilities across groups do not explain differences in marginal effects on these binary outcomes, it would imply that any observed heterogeneity is not mechanical and therefore reflects heterogeneous effects on skills and latent predispositions.