Optimal School System and Curriculum Design: Theory and Evidence*

Glenn Ellison and Parag A. Pathak†

February 2025

Abstract

This paper develops a model of education production and uses it to study optimal school system and curriculum design. Curriculum design is modeled as a time-allocation problem. A school teaches students many skills, and allocates time to different skills based on the characteristics of students. The model is used to re-interpret studies of the achievement effects of selective schools based on comparisons of admitted and non-admitted students near admissions cutoffs. A zero effect for the marginal student is consistent with a highly effective school with optimally designed curricula and need not imply ineffective schools. The intepretation depends on the alignment between skills used to measure outcomes relative to skills taught in the curriculum. The model generates several novel empirical predictions, which we examine at a prominent exam school. Results of these tests support the model and are hard to rationalize if exam schools are completely ineffective.

*We thank Josh Angrist and Eddie Lazear for helpful discussions. Ankur Chavda, Nati Mulat, Tamar Oostrom, and Rahul Singh provided excellent research assistance. Eryn Heying provided invaluable administrative support. We are grateful to partners at a large urban school district who made this study possible.

†Department of Economics, Massachusetts Institute of Technology, Cambridge MA 02142 and NBER, e-mail: gellison@mit.edu and ppathak@mit.edu

1 Introduction

A growing body of research studies how education may affect multiple dimensions of skill (see, e.g., Cunha and Heckman (2007), Jackson (2018), and Deming (2023)). In the K-12 context, studies of school effects that focus on standardized test outcomes may provide an incomplete picture of the effects of education if schools teach skills that are not measured by these assessments. For example, Abdulkadiroğlu, Angrist, and Pathak (2014) and Dobbie and Fryer (2014) use regression discontinuity (RD) designs to study elite exam schools in Boston and New York and find little benefit for marginal students on standardized tests. However, students at the Boston Latin School must take four years of Latin, and state assessments do not test Latin skills. Moreover, students at exam schools can take many more advanced courses than at traditional schools, and state assessments typically do not measure performance on this advanced material. This raises the question of how the nature of the educational process affects the interpretation of studies of selective schools.

This paper develops a theory of optimal school system and curriculum design. In the model, a school with higher ability students finds it optimal to design a curriculum targeted for these students. Curriculum design is a time-allocation problem in which a school allocates time to teach different skills based on student characteristics. We use the model to re-interpret studies of the achievement effects of selective schools. If a school has a homogenous additive effect on educational outcomes, then a zero effect on achievement at an admissions cutoff for marginal students implies the school is ineffective. However, if the division of students across schools and curriculum design are viewed as choice variables, then a zero effect at an admissions cutoff can also have a very different interpretation: the school system and curriculum have features consistent with optimal choices. The model illustrates how the effects of a school and the interpretation of estimates from discontinuity-based research designs depend on the alignment of the skills used to measure outcomes relative to other skills in the curriculum. The model also generates several novel predictions, which we empirically examine at a prominent exam school. The patterns we uncover are hard to rationalize if selective schools are completely ineffective.

The paper starts by presenting a simple model in which a school system can choose the curricula at two schools and divide students between the schools in Section 2. Curriculum design amounts to choosing the type of student that each school will serve best. This simple formulation of curriculum leads to an impossibility result about identifying the structure of education production. Based on the estimates of achievement effects for the marginal student at admissions cutoffs, it is impossible to distinguish a model of education production in which a selective school has an optimal curriculum

and provides a substantial benefit from a model of education production in which the selective school is ineffective.

This impossibility result motivates a richer analysis of the curriculum-matching to distinguish between these two explanations. We develop several predictions of the curriculum-matching model that can be examined empirically. First, in the curriculum-matching model, discontinuities in the *slope* of the outcome-admissions test relationship should be present at admissions cutoffs. Second, all students, even those with higher achievement at a less selective school, maximize their ex-ante expected achievement by stating a preference for the more selective school over the less selective one. We also explore an extension with peer effects and show that the model's predictions are sensitive to the structure of the peer effects. Conclusions about peer effects from RD studies are, therefore, joint tests of the functional form of how peers affect achievement together with the existence of peer effects given that functional form.

We next build a more explicit model of curriculum design as a time allocation problem in Section 3. In the model, schools can teach a continuum of skills and must decide how much time to devote to each skill. The optimal allocation of time to different skills would vary with the set of students that a school serves. In part, the model serves as a microfoundation for the reduced-form curriculum modeling in Section 2. The model shows that the lack of an achievement effect at an admissions cutoff depends on the alignment between the test used to measure outcomes and the set of skills a school teaches. The model predicts an upward jump at the cutoff in the difference between students' scores on tests emphasizing more and less difficult material. It also predicts discontinuities at admissions cutoff on individual test questions if performance on these questions captures different skills taught at the school.

Section 4 presents several model-motivated analyses of data on applicants to Boston's premier exam school, Boston Latin School (BLS). The data set contains information on seventeen cohorts of students who applied to the school as 6^{th} graders. The subsequent performance data include each student's score on each question on the 10^{th} grade MCAS exam, which Massachusetts public school students must pass to graduate from high school, and each participating student's total score on the PSAT and SAT exams. We start by showing that the finding in Abdulkadiroğlu, Angrist, and Pathak (2014) of no discontinuity in Grade 10 MCAS scores at the BLS cutoff remains true in our expanded dataset. We then examine several predictions of the curriculum-matching framework in the order in which they were developed in Sections 2 and 3. First, we look for evidence of discontinuities in PSAT and SAT outcomes at admissions cutoffs, and discontinuities in in the slope of the outcome-admissions test relationships.

BLS admissions increases performance on PSAT Math and SAT Math and English, highly relevant outcomes for college-bound students. Consistent with our theoretical model, it also changes the slope of the outcome-admissions test outcome for English outcomes measured by MCAS, PSAT, and SAT, and SAT Math. Next, we show that application patterns are consistent with students being influenced by expected learning and idiosyncratic preferences: the fraction of students expressing a preference for BLS over the next-highest performing school is increasing in the entrance exam score and remains above 50% even for students with scores well below the cutoff. After that, we examine question-level MCAS performance and find evidence consistent with the model's prediction that there should be discontinuities at the cutoff in student performance on many questions. Fourth, to examine the prediction that discontinuities at the margin should align with test difficulty, we report RD estimates using the difference between students' SAT and MCAS scores (and between PSAT and MCAS scores) as the dependent variable. Consistent with the model, we find substantial jumps at the cutoff in several of these variables. Finally, we consider Advanced Placement (AP) outcomes, which cover more advanced curriculum taught at BLS. BLS increases AP test-taking and performance across a range of measures. BLS admissions also increases graduation from a four-year college by 7.5 percentage points, though the estimate is marginally significant.

This paper is related to two main literatures. The first related strand of literature involves settings where curriculum may be important. This includes studies of selective schools, tracking, and gifted and talented programs such as Duflo, Dupas, and Kremer (2011) Bui, Craig, and Imberman (2014), Pop-Eleches and Urquiola (2013), Card and Guiliano (2016), and Cohodes (2020). A related literature studies the effects of attending a selective university (see, e.g., Dale and Krueger (2002) and Mountjoy and Hickman (2021)). Three studies find positive earnings effects of selective public univerisities using an RD design (Hoekstra (2009), Zimmerman (2014), and Bleemer (2024).) Chetty, Deming, and Friedman (2023) study selective private universities and consider non-traditional outcomes related to societal leadership which may be a focus of these elite institutions.[1] This work shares a focus with ours because it emphasizes the importance of considering outcomes related to the skills the educational institution targets. Furthermore, studies of accountability reforms emphasize the importance of endogenous curricula. For example, Jacob (2005) documents that Chicago's school accountability system led to achievement gains on skills emphasized on a high-stakes exam, but not on a low-stakes exam. Like our work, Jacob (2005) also uses item-level responses to measure effects on test-specific

[1] Specifically, the paper studies "non-monetary measures of upper-tail success, such as attending elite graduate schools or achieving positions of influence in public service."

skills.

Second, this paper is related to literature in education economics, which uses microeconomic models to interpret reduced form empirical facts. For example, Lazear (2001) develops a "disruption" model of education production to rationalize empirical patterns on class size effects. Urquiola and Verhoogen (2009) studies optimal class-size choices and school choice by households to show that the assumptions underlying class-size RD designs are also associated with discontinuous changes in average family income and, therefore, invalidate the underlying RD assumptions. Our work complements these efforts by examining two distinct interpretations of reduced form patterns about achievement effects of selective schools and by providing new empirical tests to adjudicate between them.

2 A Curriculum-Matching Model of School System Design

This section presents a model of school system design. The key building blocks of the model are an agreed-upon outcome ("achievement") determined by a student's learning style and the curriculum of the school to which they are assigned. The school system observes some proxy for each student's learning style and chooses both the school assignments and the school's curriculum to maximize the sum of expected achievement.

2.1 Model of School System with Selective Schools

Consider a school system consisting of two schools serving a continuum of students with learning styles θ (hereafter, their "type"), continuously distributed with full support on the interval $[\underline{\theta}, \overline{\theta}] \subset \mathbb{R}$. Without loss of generality, we can assume that θ is uniformly distributed on $[0, 1]$ by redefining θ to be the percentile of the student in the distribution. Suppose the school system observes a continuously distributed signal r_i for each student i's learning style. Assume that r is continuously distributed with full support on the interval R, that the signal is positively related to the learning style in the sense that (θ, r) have the monotone likelihood ratio property, and that the conditional density $g(\theta|r)$ is a smooth function of r.[2]

A school system design consists of two choices: 1) an assignment function $A : R \to \{1, 2\}$ giving the school to which each type of student is assigned, and 2) a choice of curricula, c_1 and c_2, for each

[2]The MLRP assumption is that $f_r(r|\theta)/f_r(r'|\theta)$ is increasing in θ when $r > r'$. It holds, for example, if $r_i = \theta_i + \eta_i$ with η_i an independent draw from a distribution that satisfies $f_\eta(x - \Delta)/f_\eta(x)$ increasing in x when $\Delta > 0$. This condition is satisfied by many common distributions, including the Normal.

school. Let the achievement of a student i of type θ_i in a school s with curriculum c_s be given by

$$y_i = h(\theta_i) + m(\theta_i, c_s) + \tau_s + \epsilon_i. \tag{1}$$

The curriculum match function, $m(\theta, c)$, reflects that students have higher achievement when the curriculum is matched to their learning style. We assume that $m(\theta, c)$ is a smooth function with increasing differences and that there does not exist a single curriculum that is optimal for all student types. For example, the quadratic function $m(\theta, c) = -(\theta - c)^2$ satisfies these conditions. We assume that $h(\theta)$, which may reflect ability, is differentiable, that τ_s, a component of education production that is the same for all students attending school s, has differences across schools that are not so large as to make it optimal to concentrate all students in the inherently better school, and that the ϵ_i are random shocks independent of all other variables.

The objective of the school system is to maximize average student achievement through its choice of curricula and student assignments. It is worth noting that the model restricts each school to offering a single curriculum and abstracts away from the costs of adopting different curricula. It also rules out capacity constraints or crowding costs since any division of students across schools is assumed possible and we have not included any enrollment-related costs.

2.2 Grouping Students by Learning Style

This **curriculum-matching model** generates a rationale for grouping students by their learning style, as summarized in the following proposition.

Proposition 1 *In any optimal school system design, there exists a cutoff $\hat{r}$ such that all students with $r_i > \hat{r}$ attend one school and all with $r_i < \hat{r}$ attend the other.*

Proof. In the optimum, c_1 cannot equal c_2 because aggregate achievement can be improved by assigning all students except those in a small interval around some type θ' for which c_1 is suboptimal to school 1 and sending students with types very close to θ' to school 2 with $c_2 = \arg\max_c m(\theta', c)$.

Suppose the optimal design has $c_1 > c_2$. Define $y(\theta, c) = h(\theta) + m(\theta, c)$. Because m has increasing differences, $y(\theta, c_1) - y(\theta, c_2)$ is increasing in θ. Given that (θ, r) satisfies MLRP, the conditional distribution, $g(\theta|r)$, is increasing in r in the FOSD sense. Hence, $\mathbb{E}_{g(\theta|r)}[y(\theta, c_1) - y(\theta, c_2)]$ is increasing in r, which implies that if it is weakly optimal to send a student with signal r to school 1, then it is strictly optimal to send all students with higher signals there. ■

2.3 Continuity of Achievement at Cutoffs

Our next result examines discontinuities in achievement at admissions cutoffs. It shows that discontinuities do not exist with optimal curriculum design.

Proposition 2 *Let $\hat{r}$ be the cutoff type in the optimal school assignment. Write $y^*(r)$ for the expected achievement level of a type r student when school assignments and curricula are chosen optimally. Then, $y^*(r)$ is continuous at $\hat{r}$.*

Proof. Define $y_s(r) \equiv \mathbb{E}_{g(\theta|r)}[y(\theta, c_s)] + \tau_s$. The continuity of h and m as functions of θ and of $g(\theta|r)$ as a function of r imply that $y_1(r)$ and $y_2(r)$ are both continuous. The distribution of r is assumed to have full support, so if $\lim_{r\to\hat{r}^-} y_2(r) < \lim_{r\to\hat{r}^+} y_1(r)$, then expected achievement could be increased by moving an interval of students with signals just below $\hat{r}$ to school 1. If the opposite inequality holds, achievement can be increased by shifting students with types just above $\hat{r}$ to school 2. ■

The argument for this result is intuitive and general. If there is a discontinuity at the cutoff, then average achievement can be improved by reassigning the students on the low side of the discontinuity to the other school. Note that the result requires that students can be assigned freely across schools. It would not necessarily hold, for example, if capacity constraints limited the number of students who could be assigned to school 1.

Proposition 2 implies that an RD estimate of the effect of attending a selective school is zero if curricula and student assignments are optimally designed. An alternate model in which this would also be true is what we refer to as the **zero value-added model** of selective schools: school effects are the same for all schools ($\tau_1 = \tau_2$) and there are no curriculum-matching effects ($m(\theta, c) = 0$ for all θ and c). We collect testable implications of the curriculum-matching model as follows:

> Prediction 0: *In both the curriculum-matching model and the zero value-added model there is no discontinuity in student achievement at the admission cutoff.*

With only a zero effect on achievement at the admissions cutoff, we cannot distinguish between two distinct models of education production with substantially different implications. In the zero value-added model, the assignment of students to schools is irrelevant for aggregate student achievement and for each individual student's achievement. In the curriculum-matching model, assignment is not

irrelevant: different assignments lead to distinct levels of individual and aggregate achievement. The identical implications of both models for changes in achievement effects at admission cutoffs do not mean that the two models cannot be distinguished from one another. In the analysis below, we highlight several testable differences that take advantage of a more comprehensive analysis of the relationships between assignments and measures of achievement.

2.4 Slope Changes at Admissions Cutoffs

Although the curriculum-matching model predicts that there should be no discontinuity in levels of student achievement at an admissions cutoff, the fact that curricula will differ across the cutoff will typically lead to other differences at the cutoff. Our next result shows that the model predicts that there will be a discrete increase in the slope of $y^*(r)$ at the cutoff.

Proposition 3 *Let $\hat{r}$ be the cutoff type in the optimal school assignment and let $y^*(r)$ be the expected achievement level of a type r student when school assignments and curricula are chosen optimally. Then, the derivative of $y^*(r)$ is discontinuous with an upward jump at $\hat{r}$.*

Proof. The change in the derivative at $\hat{r}$ is

$$\begin{aligned}
&\lim_{r\to\hat{r}^+}\frac{d}{dr}\int_\theta (h(\theta)+m(\theta,c_1))g(\theta|r)d\theta - \lim_{r\to\hat{r}^-}\frac{d}{dr}\int_\theta (h(\theta)+m(\theta,c_2))g(\theta|r)d\theta \\
&\quad = \int_\theta (h(\theta)+m(\theta,c_1))\frac{dg(\theta|r)}{dr}\bigg|_{r=\hat{r}} d\theta - \int_\theta (h(\theta)+m(\theta,c_2))\frac{dg(\theta|r)}{dr}\bigg|_{r=\hat{r}} d\theta \\
&\quad = \int_\theta (m(\theta,c_1)-m(\theta,c_2))\frac{dg(\theta|r)}{dr}\bigg|_{r=\hat{r}} d\theta \\
&\quad = \int_\theta \frac{d}{d\theta}(m(\theta,c_1)-m(\theta,c_2))\frac{d(1-G(\theta|r))}{dr}\bigg|_{r=\hat{r}} d\theta.
\end{aligned}$$

The final step follows by integration by parts since $\frac{dG(\overline{\theta}|r)}{dr}$ and $\frac{dG(\underline{\theta}|r)}{dr}$ are both zero. The first term of the product in the integral is strictly positive. The second term is non-negative and positive for at least some θ by the assumption that $\frac{dg(\theta|r)}{dr}$ is not identically zero. Hence, the integral is positive. ■

A simple intuition for why the slope changes at admissions cutoffs involves considering the case where $r=\theta$ is a perfect proxy for the learning style. In that case, student achievement varies with r in the neighborhood around $\hat{r}$ for two reasons: (1) the school-independent student performance $h(r)$

varies with r, and (2) the curriculum match $m(r, c_s(r))$ varies with r. The former has the same left and right derivative at $\hat{r}$. But the latter differs. To the left of $\hat{r}$, the curriculum match is decreasing in r because it is worse to match students to the too-low curriculum at school 2. To the right of $\hat{r}$, the curriculum match is increasing in r because the students are getting better suited to the too-high curriculum at school 1.

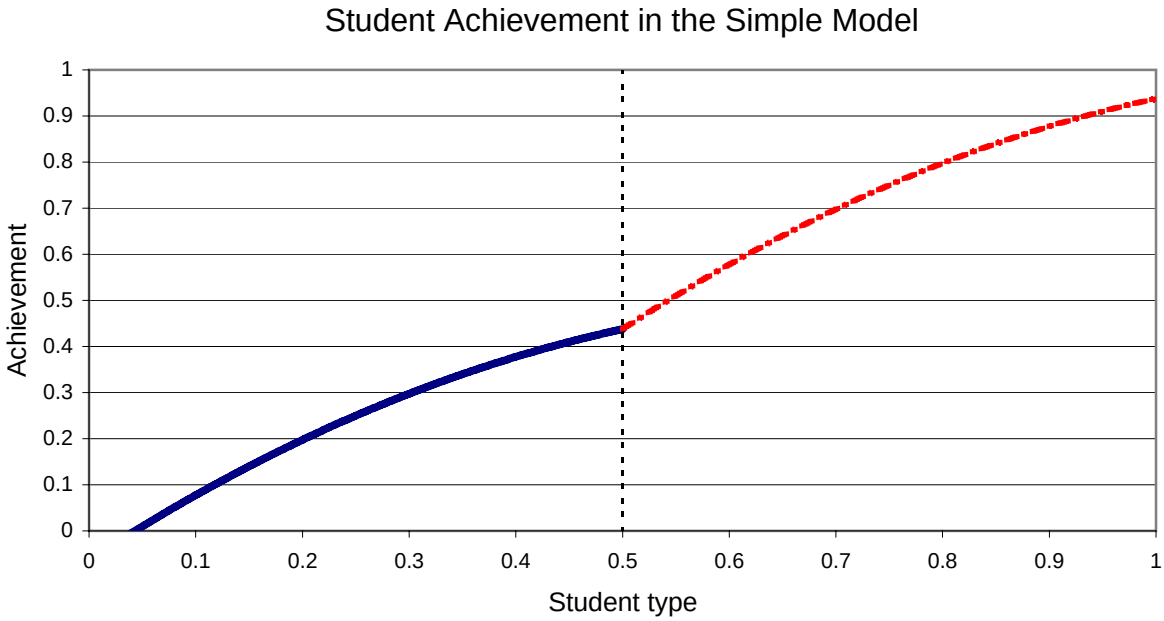
The proposition presents a formal statement about left and right derivatives at a single point $\hat{r}$, but the underlying economic intuition implies slope differences for a broader range of r values than just $\hat{r}$. To the left, mismatch is worsening. To the right, mismatch is improving. The slope will therefore be smaller on the left than on the right over a broader range, unless $\mathbb{E}_{g(\theta|r)}h(\theta)$ is highly nonlinear around $\hat{r}$.

We illustrate the change in slopes in the model with a simple calculation. Suppose that student types θ are uniformly distributed on $[0, 1]$, that $r = \theta$, and that the loss from curriculum match is quadratic: $m(\theta, c) = -(\theta - c)^2$. Then, the optimal design uses a cutoff of $\hat{\theta} = \frac{1}{2}$ and optimal curricula are $c_1 = \frac{1}{4}$ and $c_2 = \frac{3}{4}$. In the graph below, the bold blue line on the left side plots the achievement of students attending school 2. The red dashed line on the right plots the achievement of students attending school 1. The level of achievement is not discontinuous at the cutoff, but the slope of the relationship between type and achievement increases at the cutoff. Note also that due to the assumed convexity of the mismatch costs (implying concave match quality), an estimate of the slope change that wasn't fully local would *underestimate* the jump in the derivative at the cutoff that is due to the match-quality effect.

PREDICTION 1: *In the curriculum-matching model, there is an upward jump in the slope of the performance-entrance score relationship at the admissions cutoff.*

This prediction distinguishes the curriculum-matching model from that the zero value-added model. In the zero value-added model, there will be no kink in the relationship between admissions scores and student achievement at the cutoff, provided $E_{g(\theta|r)}h(\theta)$ is differentiable at $\hat{r}$.[3]

[3]A reader might wonder how well one can estimate slope changes local to cutoffs. For example, if $E_{g(\theta|r)}h(\theta)$ is convex around $\hat{r}$, a highly non-linear relationship may be mistaken for an upward jump in the slope. Convexity might arise if each percentile increase in r is associated with a larger increase in future test scores as we move further into the distribution's right tail. This would occur, for example, if r were a perfect predictor of future test scores, and these are normally distributed. On the other hand, if the entrance exam loses power to distinguish between high-achieving students, it would generate a concave relationship, which would bias against detecting a positive slope change. We suspect highly nonlinear relationships are unlikely in small windows around the cutoff in our application (which has a



2.5 Application Behavior at Admissions Cutoffs

Abdulkadiroğlu, Angrist, and Pathak (2014) note the tension between the great parent interest in sending their children to selective schools and the absence of a discontinuous jump in student outcomes around admissions cutoff. They suggest that parents mistakenly equate attractive peers with high value-added or value schools for reasons other than their impact on learning.

The curriculum-matching model offers another rationale for such parent demand. The absence of a benefit at the margin is still consistent with substantial value to attending a more selective school for students who are above the cutoff. If parents are uncertain about their child's θ and r at the time when they apply, they should express a preference for the more selective school because *conditional on the preference being relevant*, the student will have a value of r above the admissions cutoff and hence benefit from attending the more selective school.

Consider an extension of the curriculum-matching model where students receive a signal $t_i = r_i + \xi_i$ related to how the school system will assess their learning style, with ξ_i a continuously distributed random variable independent of (θ_i, r_i). Suppose that student i receives utility

$$u_i = y(\theta_i, c_s) + \nu_{is}$$

cutoff not far out in a tail).

if assigned to school s, where ν_{is} is an independent random preference shock with a density symmetric around zero. Suppose that students submit their preferences between schools 1 and 2 after observing t_i, but without knowing θ_i or r_i, to maximize their expected utility. Then, all students, even those who will likely have higher achievement in school 2, list school 1 as their first preference with probability at least one-half.

Proposition 4 *In the model above, the probability $p(r)$ that a student later classified of type r lists school 1 as his first choice is at least $\frac{1}{2}$ for all r.*

Proof. The preference order submitted by the student is irrelevant if $r_i < \hat{r}$. Conditioning on $r \geq \hat{r}$, the expected gain in achievement from school 1 is $\mathbb{E}[y^*(r)|r \geq \hat{r}]$, which is strictly positive because $y^*(r)$ is positive for all $r > \hat{r}$. With symmetricly drawn preference shocks ν_{is}, the probability that school 1 is the utility-maximizing choice is greater than one-half. ■

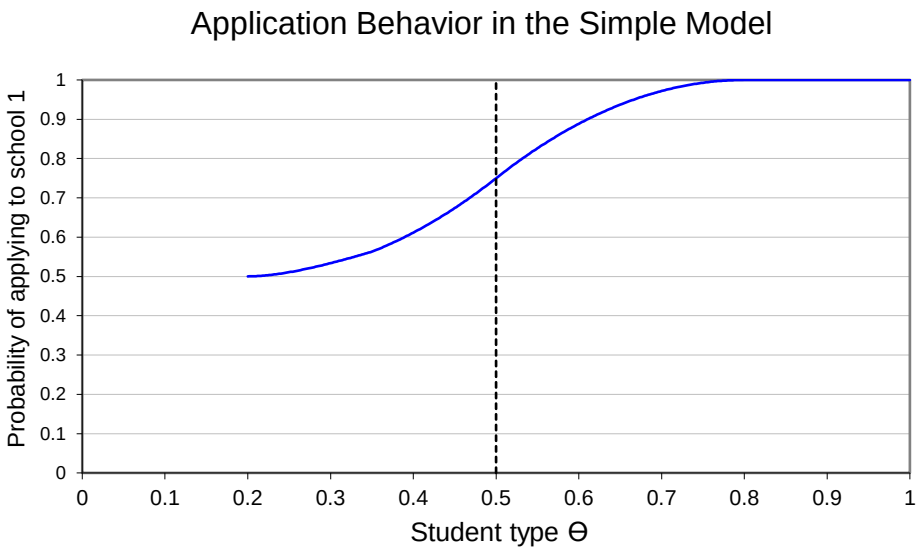
In this model, the observed probability of submitting a preference for school 1 will often increase with r. Students with higher r will be more likely to have observed higher values of t, and students who observed higher values of t will think that r is higher, and hence often have a higher expected benefit from choosing school 1. The conditions required for this comparative static, however, are not as neat as those as in the previous subsections: an FOSD increase in r will also make it more likely that the student is admitted to school 1. If the additional probability comes disproportionately from the increase in the chance that r is only slightly higher than $\hat{r}$, the excected achievement advantage conditional on being admitted to school 1 could be nonmonotonic in t, and this makes it possible to specify distributions for which the probability of stating a preference for school 1 is nonmonotonic in r.

Rather than presenting a theorem with cumbersome conditions sufficient to rule out the pathological case, we present an example which captures the main insights. Consider the case where θ is uniform on $[0, 1]$, $\tau_1 = \tau_2$, $h(\theta) = \theta$, $m(\theta, c) = -(m - c)^2$, the school system observes θ perfectly ($r_i = \theta_i$), that the noise in observing r, ξ, is distributed uniformly on $[-\sigma, \sigma]$, and the relative preference for school 1, $\nu_{i1} - \nu_{i2}$ is uniformly distributed on $[-\delta, \delta]$. Suppose also that $\sigma < 0.25$ and $\delta \geq \sigma$.

In this model, students who ex post are right at the cutoff between the two schools pick school 1 with probability strictly above one-half: $p(\hat{r}) = \frac{1}{2} + \frac{\sigma}{4\delta}$. To see this, note that students who scored exactly at the cutoff (which is $\frac{1}{2}$ in the uniform model) observed different signals before applying. On the low end, some students received the signal $\theta' = \hat{\theta} - \sigma$. These students knew that the only way

their decision could matter is if their true type is exactly $\hat{\theta}$. Therefore, they will submit the decision that they would make if their achievement is exactly equal at the two schools. Hence, exactly half of them will list each school as their first choice. Students with θ' in $(\hat{\theta}-\sigma, \hat{\theta}+\sigma)$ will condition on their true type being uniformly distributed on $[\hat{\theta}, \theta'+\sigma]$.[4] This leads them to apply to school 1 with a probability strictly greater than one-half. For example, students who saw signal $\hat{\theta}+\sigma$ apply to school 1 with probability $\frac{1}{2}+\frac{\sigma}{2\delta}$ because their expected achievement gain is σ. Expressions are linear between these extremes, which gives the formula above.

In this case, it is also possible to show that $p(r)$ is weakly increasing in r, that its derivative at $\hat{r}$ is $\frac{1}{4\delta}$, and that $p(r)=1$ for $r > \hat{r}+\sigma+\delta$. We graph the first-choice probability for school 1 computed for $\sigma = 0.15$, and $\delta = 0.15$. In this specification, the probability of listing school 1 first increases smoothly from one-half for students with very low types to one for students with very high types and is about three-quarters for students with types near the cutoff.[5]



[4] This formula assumes that $\sigma < 0.25$.

[5] The fact that the probabilities reach $\frac{1}{2}$ and 1 rather than just asymptoting to these values is a consequence of the assumption that ξ and η have finite supports. For example, students with high enough types will get a signal that implies, with probability one, that the gain from attending the best school is larger than the range of the taste shock. The finite support also results in choice probabilities being indeterminate for students whose signals are so low as to make it impossible that they will gain entry into the more selective school. For the exhibit, we have assumed that these students list school 1 first with probability one-half, i.e. they behave as if they believe that conditional on being admitted to the elite school their type is $\hat{r}$ (the lowest type consistent with admission) with probability one.

PREDICTION 2: *In the curriculum-matching model, if parents receive an imperfect signal about their children's types and have symmetrically distributed idiosyncratic preferences about schools, then the probability of ranking the top school first will be strictly greater than one-half for students at the cutoff and will be increasing in the admissions score.*

2.6 Peer Effects in Education Production

So far, the model assumes that student achievement depends on a student type θ_i, the school, and the curriculum. This section examines how results change if achievement also depends on peer effects.

Suppose that achievement is now given by:

$$y_i(\theta_i, c) = \theta_i + m(\theta_i, c) + b \sum_{j \neq i} w_{ji}\theta_j, \tag{2}$$

where $b > 0$ reflects the strength of peer effects, and $w_{ji} \geq 0$ is the influence that student j has on student i given the school assignments. In such a model, it is natural to assume that peer effects satisfy $\sum_k w_{jk} = 1$ for all j, i.e., that the allocation can affect which students are helped or harmed by student j's presence but cannot affect the total impact of student j on others. We also assume that $w_{jk} = 0$ if students j and k do not attend the same school. Given these assumptions, an immediate extension of Proposition 1 is:

Proposition 5 *When outcomes are given by equation (2), in the optimal assignment there is a cutoff $\hat{\theta}$ such that students with type above $\hat{\theta}$ are assigned to school 1. Student achievement with the optimal assignment and curricula will have an upward jump at the cutoff with the magnitude of the jump being equal to the difference in the achievement boost from peer effects received by students with types just above and below the cutoffs.*

Both Pop-Eleches and Urquiola (2013) and Abdulkadiroğlu, Angrist, and Pathak (2014) suggest interpreting reduced-form RD estimates of offers at exam schools as changing the characteristics of peers with whom a student attends school. Proposition 5 shows that the magnitude of the reduced-form discontinuity from this model depends on how peers affect achievement. A common assumption is that each student is equally influenced by all other students at the school. In this case, the jump in peer effects at the cutoff is

$$b(\mathbb{E}[\theta|\theta > \hat{\theta}] - \mathbb{E}[\theta|\theta < \hat{\theta}]).$$

At the other extreme, suppose that students interact with peers most similar to themselves. In particular, students receive peer effects from the single most similar student in the school, i.e., $w_{jk} = 1$ if student θ_j is closer to θ_k than is $\theta_{k'}$ for any $k' \notin \{k, j\}$, and $w_{jk} = 0$ otherwise.[6] In such a case, there is no discontinuity at the cutoff: the increase in peer effects at the cutoff is b times the difference between the second-lowest type above the cutoff and the second-highest type below the cutoff. This will be of the same magnitude as the difference between the values of θ of the students who are closest to the cutoff on each side, so a regression discontinuity analysis would not find any jump.

As a result, statements about peer effects from RD estimates of selective schools should be seen as joint tests of both the functional form of peers together with the existence of peer effects given that functional form. In the curriculum-matching model, it is not possible to make predictions about what jumps at cutoffs mean for peer effects without assuming a particular functional form of peer effects.

3 A Model with Explicit Curriculum Design

Up to now, our analysis has been based on a black-box approach to curriculum design: each school chooses a real number from a one-dimensional choice set, which represents the curriculum. In the example with a quadratic mismatch function, the curriculum choice can be thought of as choosing the student type that the school serves best, and the effect on other students is a function of the distance between their type and the best-served type. In this section, we develop a more explicit model of curriculum design in which schools teach many skills. Curriculum design consists of choosing different allocations of time to skills both on the extensive and intensive margin. In this expanded model, we can discuss how student performance on an assessment depends on the material assessed. The model also serves as a microfoundation for the model in the previous section.

3.1 A Model of Skill Accumulation

Consider a model in which achievement consists of learning a subset of possible skills indexed by $x \in [0, 1]$. Students learn these skills probabilistically as they are taught. In particular, if a school teaches some skill x with unit intensity in a time interval of duration dt, we assume that a type θ student who does not know the skill will learn it with probability $\theta x dt$. Students who already know skill x will receive no benefit from this instruction. Note that skills corresponding to a larger x are

[6] Note that to keep the model simple we have departed from the assumption that $\sum_k w_{jk} = 1$ for all j, allowing some students to influence 0 or 2 peers.

easier to learn and that students with a higher θ can be thought of as higher ability students who will (in expectation) learn any skill more quickly.

The instantaneous model implies that if skill x is taught for total time t, the probability that a student of type θ learns the skill by the end of the period is

$$a(x; t, \theta) = 1 - e^{-\theta x t}.$$

While the model is stated as one in which learning a skill is a binary event, it can also be interpreted as one in which a student's mastery of the skill increases smoothly in the time devoted to the skill with this functional form. Decreasing returns to the time spent teaching any skill is implicit in this formulation.

3.2 Curriculum Design as a Time-Allocation Problem

View curriculum design as a time-allocation problem. Choosing a curriculum consists of choosing a function $t(x)$ describing how much time will be spent teaching each skill x. We define the achievement of a student of type θ exposed to curriculum t as a weighted average of the student's learning of each skill, with the weights $v(x)$ reflecting the importance assigned to skill x. Given a curriculum $t(x)$, expected achievement is:

$$y(\theta, t(\cdot)) \equiv \int_0^1 v(x) \left(1 - e^{-\theta x t(x)}\right) dx.$$

With these definitions, the model provides a microfoundation of the model of the previous section but with curriculum consisting of multiple dimensions. For each type θ, there is an optimal curriculum, and achievement decreases as the curriculum differs from the optimum.

We first consider a single school that serves a set of students with types θ distributed with density $g(\theta)$ over set Θ. Assuming that the school has total time T available for instruction, the problem of maximizing aggregate achievement is:

$$\begin{aligned} &\max_{t(\cdot)} \int_\Theta y(\theta, t(\cdot)) g(\theta) d\theta \\ &\text{s.t.} \int_0^1 t(x) dx = T. \end{aligned} \tag{3}$$

Given a curriculum design $t(x)$, the marginal value of spending additional time teaching a skill x

is

$$\mathbb{E}_g[\theta x e^{-\theta x t(x)} v(x)].$$

Initially, when $t(x) = 0$ the marginal benefit of the first instant teaching skill x is simply

$$xv(x)\mathbb{E}_g[\theta].$$

The marginal benefit decreases in $t(x)$ and declines to zero in the $t(x) \to \infty$ limit. The structure of declining marginal benefits allows us to establish the following properties about the optimal curriculum design $t^*(x)$.

Proposition 6 *Let $t^*(x; v, g, T)$ be the solution to the optimal curriculum design problem (3). Then,*

(a) there exists a cutoff $\underline{w}(v, g, T) > 0$ such that the set of skills that are taught for a nonzero period of time is $\{x | xv(x) > \underline{w}(v, g, T)\}$,

(b) the lower bound $\underline{w}$ is decreasing in T, and

(c) the function $t^*(x)$ may be non-monotone in x even when $v(x)$ is constant.

The argument for (a) is that the optimal curriculum design equalizes the marginal return of teaching each skill that is taught. The marginal return to the first instant is $xv(x)\mathbb{E}_g(\theta)$. Hence, the set of skills taught must be where this expression is above some cutoff. For (b), the cutoff decreases in T because the marginal value of instructional time is decreasing. The potential non-monotonicity of $t^*(x)$ in x is a natural consequence of the decreasing returns to teaching a given skill.

Suppose $v(x) = \overline{v}$ for all v. At the lower bound of the set of skills taught, i.e., for $x = \underline{w}/\overline{v}$, $t^*(x)$ is zero. The t function is increasing in x for slightly larger x since there is a greater benefit to teaching a skill that will be learned more quickly. But when $t(x)$ is larger, another effect can dominate: the easier skill is more likely to already have been learned, making additional time teaching it less valuable. The numerical example in the next subsection illustrates this possibility.

3.3 Example of Optimal Time Allocation to Teaching Different Skills

In this subsection, we provide a closed-form expression for $t^*(x)$ when a school places a constant value $v(x) = 1$ on each skill and serves a homogeneous population of students. We evaluate these

numerically to illustrate how design varies with the ability of the student population and the time available for instruction.

If all students have type θ, then the condition that the marginal benefits are equalized across skills implies that

$$\theta x e^{-\theta x t^*(x)} = \lambda,$$

for all x with $t^*(x) > 0$. Re-arranging this expression, we have

$$t^*(x) = \frac{\log(\theta x) - \log(\lambda)}{\theta x}.$$

The time spent on skill x is monotone increasing if $\lambda > \theta/e$. Otherwise, it is non-monotone with a maximum at $\frac{e\lambda}{\theta}$. Note that λ is equal to θ in the $T \to 0$ limit and decreases to 0 as $T \to \infty$. Hence, the $t^*(x)$ function is increasing in x when T is small and becomes non-monotone when T is larger.

To compute the optimal time spent teaching skill x, we solve for λ using the fact that the total instructional time equals T:

$$T = \int_{\lambda/\theta}^{1} \frac{\log(\theta x) - \log(\lambda)}{\theta x} dx.$$

Integrating by parts gives

$$T = \frac{1}{2} \frac{(\log(\theta) - \log(\lambda))^2}{\theta},$$

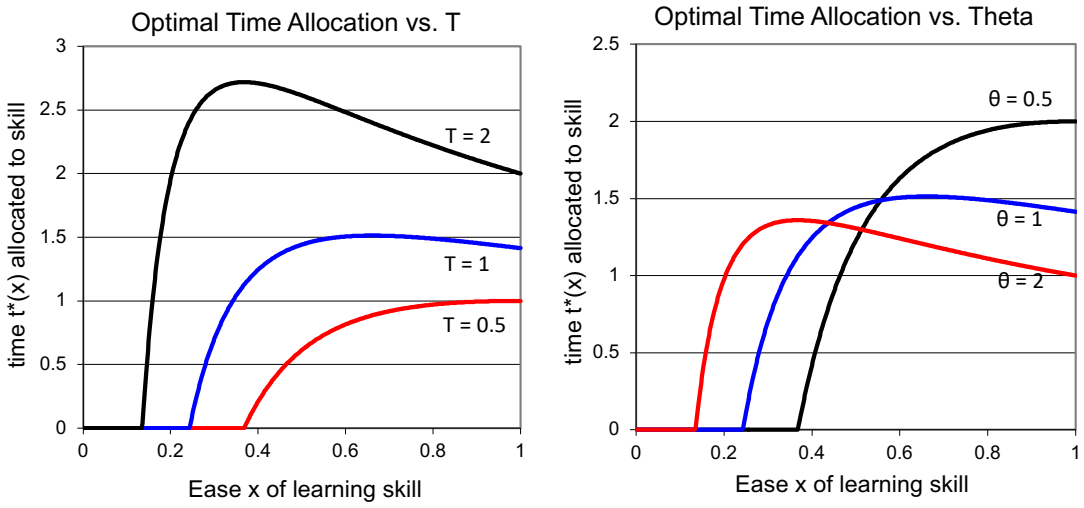
which implies that $\lambda = \theta e^{-\sqrt{2\theta T}}$. Hence, the optimal solution is:

$$t^*(x) = \frac{\log(x) + \sqrt{2\theta T}}{\theta x} \quad \text{for } x \in \left[e^{-\sqrt{2\theta T}}, 1\right].$$

We illustrate the shape of $t^*(x)$ compared to T and type graphically below. In the first panel, the lowest curve is for $T = 0.5$. It is monotone increasing. The two higher curves are for $T = 1$ and $T = 2$. As T increases, the optimal curriculum teachers a broader set of skills and the time allocations become non-monotone.

The second panel shows how the optimal time allocation varies with θ when T is fixed at one. The rightmost curve is the optimal time allocation for students with $\theta = 0.5$. It is monotone increasing. The other two curves are the optima for students with $\theta = 1$ and $\theta = 2$. They extend to lower x skills and spend progressively less time on the high-x activities and more on the low-x activities. The contrast between the curricula is particularly striking at the high and low ends. The curriculum designed for $\theta = 2$ students spends only about half as much time on the easiest skills (those with x near one) as

the curriculum designed for $\theta = 0.5$ students. The skills to which the optimal curriculum allocates the most time are right around the cutoff of being considered too difficult to be worth spending any time at all on in the optimal curriculum for $\theta = 0.5$ students.



3.4 Explicit Curriculum Design and Student Achievement

Consider now a school system with two schools and suppose that school 1 serves student types $\theta \in [\hat{\theta}, 1]$ and school 2 serves students types $\theta \in [0, \hat{\theta}]$. Write $g_i(\theta)$ for the distribution of types in school i. Our first observation is that the allocation of time will have a *single-crossing property*: school 2 will spend more time on skills where $xv(x)$ is large and school 1 will spend more time on skills with $xv(x)$ small.

Proposition 7 *Suppose that $v(x)$ is continuous. Let $t_1^*(x)$ and $t_2^*(x)$ be the optimal time allocations in the explicit curriculum design model. Then there exists a value for w such that $t_2^*(x) > t_1^*(x)$ for all x with $xv(x) > w$ and $t_1^*(x) \geq t_2^*(x)$ for all x with $xv(x) < w$, with the inequality in time allocation being strict except when $t_1^*(x) = t_2^*(x) = 0$.*

When all skills are equally valuable, Proposition 7 states that the school serving low-type students spends more time on the basics. Decreasing returns set in more quickly at the school serving high-type students, so it limits the time on the most basic skills and instead allocates time to some harder-to-learn skills. The proposition shows that the proper generalization of this intuition to a model where

skills can also differ in value involves ordering the skills in terms of $xv(x)$ (which is the marginal return of the first instant spent on a skill). The full proof is in Appendix A.

Except in a non-generic specification where there is a positive measure of skills x with $xv(x) = w$, a corollary of Proposition 7 is that the curricula will satisfy $t_1(x) \neq t_2(x)$ for almost all x with $t_1(x) > 0$. Therefore, a full description of what every student (including those with type $\hat{\theta}$) learns depends on the student's assignment. In particular, students will learn all skills with $xv(x) < w$ better if they are assigned to school 1 and will learn skills with $xv(x) > w$ better if they are assigned to school 2. This generates our next testable prediction.

PREDICTION 3: *In the curriculum-matching model with explicit curriculum design, there will be a discontinuity in student understanding of skill x at the cutoff $\hat{\theta}$ for almost all skills x.*

RD estimates of the effects of selective schools on student achievement measure achievement using exams that comprise a large number of questions assessing a range of skills. Although students will learn each particular skill better at one school or the other, our impossibility result, Proposition 1, extends to the model with explicit curriculum design if the test students are given is aligned with the school system's learning objectives in a specific way.

Proposition 8 *Suppose students are assessed on a test that includes a mass $q(x)$ of questions on skill x. If $q(x)$ is directly proportional to $v(x)$, then there will be no discontinuity in student performance on the test at the cutoff type $\hat{\theta}$.*

The argument for this proposition is straightforward. Each student's expected score on the test is proportional to the outcome the school system is designed to maximize. As in the simple model, the lack of a discontinuity at the admissions cutoff is a necessary condition for optimality. We can, therefore, extend the earlier prediction to include the analysis of peer effects and explicit curriculum design as follows:

PREDICTION 0′: *(a) In both the zero value-added model and the optimal curriculum matching model with a test tailored to the learning objectives, there is no discontinuity in student achievement at the admission cutoff if there are no peer effects. (b) Peer effects can, but need not, lead to upward jumps at the cutoff. (c) Mismatches between the test and learning objectives can lean to upward or downward jumps at the cutoff in the curriculum design model.*

In practice, students are often evaluated on multiple tests over the course of their school career. The model above also provides predictions about where they will do relatively well and poorly. Roughly, students at school 1 will do relatively poorly on tests emphasizing basic material and relatively well on tests emphasizing difficult material. We can state this result more formally as follows.

Proposition 9 *Let $q_j(x)$ be the fraction of questions on test j testing skill x. Write $\Delta(\theta) \equiv y_A(\theta) - y_B(\theta)$ for the difference between a student of type θ's score on test A and their score on test B. Suppose the school assignment is optimal and $q_A(x) < q_B(x)$ if and only $xv(x) < w$ where w is the single crossing point defined in Proposition 7. Then $E(\Delta|\theta)$ will have an upward jump at $\theta = \hat{\theta}$.*

This proposition leads to our last prediction.

PREDICTION 4: *If more and less difficult tests satisfy the condition in Proposition 9, then the difference between students' scores on more and less difficult tests will show an upward jump at the cutoff $\hat{\theta}$.*

This prediction also distinguishes the curriculum-matching model from the zero value-added model. If scores on tests $j \in \{A, B\}$ are given by $y_{ij} = h_j(\theta_i) + \tau_s + \epsilon_{ij}$, then $\mathbb{E}[y_{iA} - y_{iB}|\theta] = h_A(\theta) - h_B(\theta)$, which has no jump at the cutoff if performance on each test is a continuous function of the student's type. To empirically examine this prediction, we will analyze the relationship between entrance exam scores and the difference between SAT (or PSAT) scores and scores on state standardized tests. The latter are a high school graduation requirement, and have a higher concentration of questions on less advanced material than the former.

4 Evidence from the Boston Latin School

This section considers applicants to the Boston Latin School (BLS) to examine predictions of the curriculum-matching model compared to the zero value-added model.

Boston Public Schools (BPS) is an appealing environment for our investigation for two main reasons: (1) BPS has a sharp admissions cutoff at its three exam schools, and admissions criteria did not change during our study period[7]; and (2) BPS has shared 17 years of student-level data that includes application preferences, entrance exam scores, and a set of student outcomes. During our

[7]BPS adopted a new exam school admissions policy in 2021, that considers the applicant's neighborhood characteristics as well as GPA (Barry 2021).

study period, admissions is based on a combination of performance on Independent Schools Entrance Exam (ISEE) and grades; throughout the paper, we refer to this combination as the entrance exam or admissions test. Within the BPS system, we focus on BLS, the most selective of Boston's exam schools, because curriculum-matching considerations are likely most salient for this selected group of students. BLS is seen as a gateway to elite post-secondary institutions and BLS students are expected to engage with rigorous college preparatory curricula that are different from curricula that would maximize MCAS scores.[8] Our analyses of student educational outcomes consider the 10th grade MCAS, 11th grade PSAT, and SAT tests.[9] We focus on these outcomes because students will have been exposed to the BLS curriculum for several years before they are assessed, the tests cover different material, and these three assessments are taken roughly in the last years of high school (between grades 10-12). We also consider effects on college enrollment using data from the National Student Clearinghouse.

Our data set overlaps with one studied in Abdulkadiroğlu, Angrist, and Pathak (2014), but contains an additional seven years of data. With this additional data, we can both look for harder-to-identify patterns than were studied in that paper, e.g. looking for changes in slopes and question-level differences or estimates on outcomes like AP and college-going, and narrow our focus to the setting where we think the curriculum-matching model is most salient, as described above. An additional benefit of the narrower focus is that all of our RD estimates are focus exclusively on applicants for 7th grade admission who list BLS as their first choice. Abdulkadiroğlu, Angrist, and Pathak (2014) reported pooled estimates from tests given at different points in time to multiple groups of students who had expressed different preferences at the time of their initial application. These pooled estimates represent a weighted average of potentially heterogeneous effects.[10] Our sample of 7th grade applicants who list BLS as their first choice contains the majority of students in their study during the overlapping years. The data appendix contains additional details on data processing.

Table 1 reports descriptive statistics on the sample. As in most large urban districts, the majority of Boston 6$^{\text{th}}$ graders are Black or Hispanic. Boston students' average scores are below state averages on MCAS assessments, and below national averages on the PSAT and SAT. 56% of students attend

[8]The BLS student handbook states (BLS 2016): "Boston Latin School is open primarily to students who intend to go to college and wish to prepare in the liberal arts tradition. Students, who are admitted only into grades seven and nine, pursue a six-year or four-year college preparatory program. The curriculum of Boston Latin School is diverse and demanding. Besides classroom work, students are expected to do about three hours of home study every day."

[9]Some students take the PSAT and SAT multiple times. For PSAT, we take the 11th grade PSAT. For the SAT, we take the maximum observed score.

[10]Abdulkadiroğlu, Angrist, and Pathak (2014) also look for discontinuities at the cutoffs for the second- and third-most selective exam schools, examine 7th and 8th grade MCAS scores in addition to 10th grade scores, and study both students who applied for 7th and 9th grade admission.

college, and 48% enroll for four or more semesters, our measure of college persistence. Sixth grade BLS applicants and the subset who listed BLS first are less likely to be minority, score higher on MCAS tests, and are more likely to attend college. Applicants who are within 20 percentile ranks of the BLS admissions cutoff are a further selected group. Fewer than half of these applicants are Black or Hispanic and the average baseline Math and English MCAS scores are about 0.7σ and 0.4σ higher than the state average. Moreover, these applicants score well above national averages on the SAT and their college attendance, persistence, and graduation rates are notably higher than the overall Boston population. These differences indicate the importance of considering measures of educational performance that are most relevant for this selected student population.

4.0 Prediction 0: Discontinuities in Standardized Test Scores at the Admissions Cutoff

We begin our investigation by reproducing a primary result of Abdulkadiroğlu, Angrist, and Pathak (2014) on our augmented sample. Specifically, we examine whether there are discontinuities at the BLS cutoff in scores on the Math and English sections of the 10th grade MCAS exam. The 10th grade MCAS exam is given to all Massachusetts public school students. It was a high school graduation requirement and plays a central role in the state's school accountability system. We normalize MCAS scores in each year to have mean zero and standard deviation one within all test-takers in Massachusetts, and scores are reported in standard deviation units (hereafter, σ). Figure 1 contains plots of 10th grade Math and English MCAS scores as a function of students' percentile rank on the admissions test among exam-school applicants in their year.

Let r_i be BLS applicant i's admissions test score, normalized so that $r_i = 0$ at the BLS cutoff. Applicants within 20 percentiles of the BLS cutoff are grouped into one-percentile bins. The dots give the average normalized MCAS scores for students in each bin. The curves are local linear regression estimates of the conditional mean function.[11]

The dashed vertical lines in the figures are the cutoffs for admission to BLS. As in Abdulkadiroğlu, Angrist, and Pathak (2014), discontinuities do not appear at the cutoff in either math scores (shown on the left) or English scores (on the right). As we've discussed, this reduced form fact is consistent with the zero value-added model and with optimal assignments in some versions of the curriculum-matching model. In the curriculum-matching model, a zero effect would occur, for example, if the school system

[11]The estimates use the edge kernel with the IK optimal bandwidth as in Abdulkadiroğlu, Angrist, and Pathak (2014).

(curricula and student assignments) are designed to maximize performance on the MCAS and there are no peer effects. It would also be zero if the curricula aim to maximize learning of material that is more difficult than what is tested on the MCAS and peer effects are of the appropriate magnitude/degree of localization to offset the downward jump that would otherwise be present, as described in Section 2.6.

4.1 Prediction 1: Changes in Slopes at the Admissions Cutoff

We turn now to the additional predictions of the curriculum-matching model. Our first exercise is motivated by Prediction 1. We examine whether there are changes in the *slope* of the achievement-entrance exam score relationship at the BLS cutoff. Recall the intuition for this result is that match quality is decreasing to the left of the cutoff (students are farther above the type the school is designed to serve best) and increasing to the right of it (students are less far below the type the higher school is designed to serve best). The shift from decreasing to increasing match quality leads to a discrete upward jump in the derivative of the achievement-entrance score relationship.

Figure 2 presents RD plots similar to those in Figure 1 for four different tests geared for post-secondary education: the PSAT and SAT math and the PSAT reading and SAT critical reading exams. The dots in the figures are conditional means for students in each one-percentile bin within 20 percentiles of the BLS cutoff. The curves are conditional means estimated via local linear regressions on each side of the cutoff. The series appear to be fit fairly well by the estimated relationships in a neighborhood of the cutoff. The reading tests in the right panel both seem to have visible (and substantial) increases in the slope at the BLS cutoff. The PSAT and SAT math figures suggest positive but smaller slope changes on those tests, although the PSAT figure shows those data to be quite noisy. Looking back at the 10th grade MCAS data in Figure 1, it was somewhat similar with a visible (but smaller) increase in the slope of the English scores at the cutoff, but no visible slope change for math.

To assess the statistical significance of the slope changes in the figure, Table 2 presents RD estimates of slope changes at the cutoffs with standard errors. Let y_{it} be the score on a given test for a student i who applied for BLS in year t. We report local-linear estimates of

$$y_{it} = \alpha_t + \gamma_0 f(r_i) + \gamma_1 D_i g(r_i) + \rho D_i + \eta_{it}, \tag{4}$$

where D_i indicates an offer of admissions (i.e. $r_i \geq 0$), $f(r_i)$ and $g(r_i)$ are potentially nonlinear

functions of the entrance exam score normalized to have slope one at $r = 0$, and the coefficient of interest is the change in slope, γ_1, at the admissions cutoff. The α_t terms control for the application year. These specifications use the optimal Imbens and Kalyanaraman (2012) (IK) bandwidth with a tent-shaped edge kernel.[12] This specification is a simplified version of the one in Abdulkadiroğlu, Angrist, and Pathak (2014) focused on a single school for 7th grade applicants who have listed BLS first.[13] Table 2 reports estimates for six different test outcomes: math and reading scores on the MCAS/PSAT/SAT.[14]

The fact that several of the estimated slope changes at the cutoff are positive and significant provides support for Prediction 1 of the curriculum-matching model. The estimated slope changes for the SAT Math, MCAS English, and PSAT English are significant at the 5% level, and that for the SAT reading is significant at the 10% level. Given the inherent difficulty in estimating slope changes in an RD framework, we see these precisely estimated slope change estimates as surprisingly strong evidence for curriculum matching.[15]

The standard model of peer effects, in which each student receives an additive benefit reflecting average student quality that accrues equally to all students in a school, would not lead to any change in slope at the cutoff. The alternate model we described in equation (2), in which each student receives a benefit proportional to the quality of the student most similar to them, doesn't lead to a discontinuity at the cutoff, nor does it produce a slope change at the cutoff. Other specifications of peer effects, however, could produce a slope change at the cutoff. For example, suppose the benefits a student of type θ receives at school s from peers increases in the average peer quality but decreases with dissimilarity from one's peers. One functional form with this characteristic is $\alpha_0\overline{\theta}_s - \alpha_1(\theta - \overline{\theta}_s)^2$.

[12] It is worth noting that our problem is conceptually different from that in studies of the regression kink design (see, e.g., Card, Lee, Pei, and Weber (2015) and Ganong and Jäger (2018)) where the probability of assignment is a kinked function of the running variable. Here, the assignment is discontinuous at the admissions cutoff.

[13] See page 154 of Abdulkadiroğlu, Angrist, and Pathak (2014). Since we focus on first-choice applicants, we do not need additional controls for assignment risk in the specification Abdulkadiroğlu, Angrist, Narita, and Pathak (2022).

[14] Appendix Table A1 reports on outcome follow-up differentials and shows no follow up difference for the PSAT, SAT, the PSAT-MCAS, or SAT-MCAS outcomes. Differences in outcome availability, therefore, seem unlikely to affect estimates for these outcomes. The one notable follow up differential is for the Grade 10 MCAS outcomes, where applicants who clear cutoffs are about 10% more likely to have an observed MCAS outcome. Results for MCAS outcomes should be treated with more caution given this differential.

[15] Note that our results are derived from local linear regressions. Standard errors of the slope changes estimated via Calonico, Cattaneo, and Titiunik (2014) would be several times larger than the effects we estimate. We feel, however, that our procedure is more appropriate for two reasons. First, under the zero value-added null hypothesis, the second derivative of $\mathbb{E}(y|r)$ is the same to the left and right of the cutoff, so our theory implies that we should impose that the bias that their method estimates and accounts for is zero. Second, given that the model suggests a positive difference between slopes to the left and right of the cutoff (not just a kink at the cutoff), we are not troubled by the informal concern that the IK bandwidths are considering a non-trivial interval around the cutoff.

This specification would make the model very similar to our simple curriculum-matching model and can produce a slope increase at the cutoff via the same mechanism.

4.2 Prediction 2: Stated Applicant Preferences

In Section 2.5, we showed that a curriculum-matching model provides a natural way to reconcile strong parental interest in selective schools with the difficulty in identifying benefits for marginal students. As summarized in Prediction 2, stated preferences for a selective school are only relevant if the student turns out to be above the cutoff, in which case benefits do exist. Hence, if purely idiosyncratic preferences are symmetric, the majority of students with any admissions test score (even those actually better matched to the less selective school) should express a preference for the more selective school. Moreover, the fraction of students at each entrance score level who prefer the more selective school should be increasing in the admissions test score level.

To examine this prediction, we report a nonparametric estimate of the probability of ranking BLS above Boston's second-most selective exam school, Boston Latin Academy (BLA), as a function of the admissions test score, controlling for the distance between the applicant's home and each school. We estimate this relationship on the full population of students who applied to either BLS or BLA. The analysis, therefore, includes many students with admissions test scores well below the admissions cutoff whose outcomes have not been shown in prior figures focusing on a neighborhood of the BLS cutoff, which is shown as the dashed vertical line. Let p_{it} equal 1 if applicant i in cohort t ranked BLS over BLA. A student is treated as ranking BLS above BLA if the student ranked both schools and put BLS higher or if they included BLS, but not BLA on their preference list. Specifically, we estimate:

$$p_{it} = \mu_t + \sum_{s \in \{BLS, BLA\}} \beta^s d_{it}^s + g(r_i) + \upsilon_{it},$$

where μ_t is a dummy variable for each cohort, d_{it}^s is the distance between the applicant's home address and school $s \in \{BLS, BLA\}$ with associated coefficient β^s, and $g(\cdot)$ is a flexible function of the admission test.[16] We control for distance because a student's distance to BLS or BLS may affect their demand, but the patterns are nearly identical regardless of whether we include these controls.

Figure 3 plots the local-linear fit of this equation. The fraction of applicants expressing a preference for BLS over BLA is above 50% for all students, even those with very low admissions test scores. The

[16] Appendix B provides more details on the distance calculation.

fraction is around 75% for students with scores near the BLS cutoff and increasing for a wide range of scores around this cutoff. The fraction is above 90% for students with the highest admissions test scores. These features are all consistent with a curriculum-matching model in which students have an imperfect signal of their match quality: those with scores around the cutoff (correctly at the time of application but incorrectly ex post) anticipate that their match quality will be substantially better at the top school *conditional on being admitted to it*, and students with the highest scores (correctly) anticipate having an even higher match quality differential.

As noted in the theory section, it is difficult to reconcile these strong preferences for the top school with a fully rational zero value-added model. Abdulkadiroğlu, Angrist, and Pathak (2014) advocate a mistake in parental perceptions of value-added, known as an "elite illusion." This explanation requires that belief errors are larger in students who score higher on the admissions exam or that there are unobserved student characteristics correlated with both admissions test scores and preferences for BLS relative to BLA.

4.3 Prediction 3: Discontinuities in Question-Specific Performance

The curriculum-matching model with explicit curriculum design suggests that there will be discontinuities in student performance on individual questions at admissions cutoff. The reason is that when a school optimizes its curriculum, the school spends different amounts of time teaching each skill. Large differences in time can exist between easy-to-master skills, which the most selective school covers quickly, and hard-to-master skills, which the least selective school omits due to time constraints. These differences will translate into different performance on particular questions if the questions measure different skills. Question-level discontinuities should exist even when there are no aggregate score discontinuities, e.g. if the test weights skills with the same importance that a school used in setting its curriculum.

We turn to an investigation of question-level data on student performance on the MCAS exams. Several factors make the estimation of question-specific discontinuities more difficult than estimating discontinuities in a test score. First, since questions change each year, the number of observations in a question-specific analysis is small. Second, an individual's score on a single question, as a 0-1 variable, is a noisy measure of skill performance; Third, question-level data is only available for the MCAS and not for the PSAT/SAT. Finally, given the (low) level of difficulty of the MCAS exam and the (high) level of achievement of students who are around the BLS cutoff, many students around the

BLS cutoff have already learned the material that some 10th MCAS questions cover as 6th graders and/or would be able to get the questions right as 10^{th} graders even if they had not been taught the material. Nonetheless, we are able to obtain some results and the evidence we find supports the curriculum-matching model.

We start with the most basic aggregate question: are there discontinuities in student performance at the BLS cutoff on at least some questions? The answer is clearly yes. Using the same estimation framework as above, we estimated 1,054 separate RD analyses: one for each 10th grade MCAS question.[17] Figure 4 presents histograms of the p-values for the significance of the estimated discontinuities, reporting separately questions on the English and Math sections of the 10th grade MCAS exam. Without any discontinuities, these histograms should be roughly uniform on $[0, 1]$, e.g., 5% of the coefficient estimates should have t-statistics indicating significance at the 5% level. Instead, we find that 16% of the math questions and 12% of the English questions have discontinuities significant at the 5% level. There is also consistently a deficit in p-values above 0.5. On the math exams, only 35% questions have such p-values. These results show question-level discontinuities are present across questions on a given test and a nontrivial fraction of questions in each subject at each grade level.

We next examine test questions for which jumps are the largest to investigate the potential role of curriculum design, exploiting knowledge of math curricula developed by one of the authors in the course of writing math textbooks, Ellison (2010) and Ellison (2013). Figure 5 shows the five 10th grade MCAS math questions with the most statistically significant downward jumps at the BLS cutoff, i.e. questions where BLS students do worse.[18] The three questions in the upper left part of the figure strike us as having something in common: they are on topics which students who have recently been studying middle school material might have an advantage over students who have spent the last two years working through a more advanced high-school curriculum. 2003 Q10 is a middle school probability problem. 2007 Q28 is best answered using "proportional reasoning" taught in middle school: when the radius is made 3 times larger and the height unchanged, the volume increases by 3^2. Whereas a high school student's first inclination may be to translate 2007 Q30 to an algebra problem, "Find $2x + 1$ if $x(2x + 1) = 36$?," finishing this solution requires factoring a quadratic with a leading term greater than one.[19] A pre-algebra approach in which the student tries out four choices is both quicker and less likely to result in a mistake. The other two questions also strike us as having

[17] These models have 0-1 outcomes corresponding to each question rather than standardized scores.

[18] Most have point estimates of about 30 percentage points on the downward jump and each is significant at the 99.9% level, setting aside multiple testing.

[19] Some students who factor correctly may make the mistake of forgetting to report $2x + 1$ instead of x.

something in common: they are on somewhat esoteric topics that one can easily see schools deciding to leave out of their curricula. 2010 Q34 seems an odd choice for a graduation-requirement test in asking when a slope is "undefined." It seems plausible that most schools would focus much more on the central skill of writing lines in slope-intercept form and using the equations this produces to solve problems and/or not use the word "undefined". A student in a rigorous algebra class might have been taught they could solve 2006 Q1 as a three-equations in three unknowns problem, but this will be tedious. Solving is super quick and unlikely to result in an error if students have been shown that the differences between adjacent terms in a quadratic sequence are linear, but we assume that only some schools have decided to include this particular method for dealing with such sequences in their algebra curriculum.[20] Absent this, students still might do better to eschew algebra and just play around with adding or subtracting numbers from each term–subtracting 5 is magic here–until they see a pattern. In summary, it seems plausible that all five questions reflect skills BLS had chosen not to emphasize in their 9th and 10th grade math classes.

Figure 6 shows the five 10th grade MCAS math questions with the most statistically significant upward jumps at the BLS cutoff, i.e. questions on which BLS students did better.[21] These have a bit less in common. The two in the upper left, 2014 Q11 and 2014 Q13, are straightforward questions that students with good algebra training should not miss, although we could imagine that some schools avoid using numbers like 3 and 9 as exponents and some students get tripped in Q13 by the negative numbers. The middle problem, which asks students to give an equation for an approximate "best fit" line for a scatterplot, is something that may not have been covered in traditional high school curricula before it was included in the 2000 10th grade Massachusetts state standards, and one could imagine that as of 2003 the math teachers at some of the schools attended by students who fell just short of BLS cutoff had not yet redesigned their curricula to include it. The other two questions, one on rounding off to the nearest whole number and one on similar triangles, are not covering rigorous high school material (although high school and middle school geometry classes both cover similarity). The rounding question seems hard to place in any grade-level math curriculum and is perhaps more likely

[20] The elegant solution here would note that the differences between adjacent terms are 3, 5, 7, and 9, so the next difference is 11, implying the answer is 41. The technique is given less than a page in *Hard Math for Middle School* and is not covered in some standard algebra textbooks. The algebra approach every algebra course would cover (not necessarily in this application) uses the fact that the nth term is of the form an^2+bn+c to solve the problem by solving $a \cdot 1^2 + b \cdot 1 + c = 6$, $a \cdot 2^2 + b \cdot 2 + c = 9$, $a \cdot 3^2 + b \cdot 3 + c = 14$. The numbers were chosen so that the solutions are small integers, $(a, b, c) = (1, 0, 5)$, and the algebra is easier than it often would be.

[21] Most have point estimates of about 30 percentage points on the upward jump, and each is significant at the 99.9% level if one ignores the multiple hypothesis testing issue.

to have been covered in science classes. While it's harder to form a big-picture takeaway, performance on these questions is consistent with BLS students learning more advanced algebra and geometry, and this skill accumulation contributes to some of the upward jumps.

4.4 Prediction 4: Discontinuities in Score Differences on More and Less Difficult Tests

Prediction 4 highlights an alternate way to test whether schools are choosing to allocate time in the curriculum considering skills that would most benefit their students. If we form the difference between student scores on a "hard" and "easy" test and examine the relationship between this difference and entrance exam scores, we should see an upward jump in the score difference variable at the admissions cutoff.

The design of the 10th grade MCAS Mathematics exam is particularly fortuitous in this regard in that it mostly tests below-grade-level material. Appendix A.2 provides some evidence on this, noting, for example, that a document documenting the 2013 exam's alignment with Massachusetts's 2011 curriculum framework maps 23 of the 42 questions to middle school standards: 5 questions are mapped to grade six standards, 13 to grade seven standards, and 5 to grade eight standards. The 10th grade MCAS English exam does not seem similarly below grade level: each test has a diverse mix of readings that includes passages from items from the high school English course canon, e.g. translations of Greek and Latin works, Shakespeare, *Beowulf*, Austin, Cervantes, Conrad, Dickens, García Márquez, Kafka, Shelley, that are ten times as long as any passage on the PSAT alongside some less challenging popular press articles and passages from non-classic books.[22] It is harder to quantify grade levels on an English test, but the books do seem to be roughly around a 10th grade level, with the non-book readings easier. Appendix A.2 provides more details.

We take advantage of differences in tests by forming four new dependent variables for each student: the difference between a student's highest score on the math/critical reading SAT and their score on the 10th grade math/reading MCAS and the difference between a student's 11th grade score on the math/reading PSAT and the 10th grade math/English MCAS. The PSAT has the advantage of being taken only five months after the MCAS. The SAT provides a larger difference in the difficulty of the material compared to the MCAS, particularly for the English exams. Subtracting scores is appropriate

[22]The MCAS tests include a few footnotes giving definitions of some words in the passages. For example, the Garcia Marquez passage gives definitions for *taciturn*, *stigma*, *breviary*, and *vignettes*.

since our normalization puts the scores on a common scale.[23]

Estimates from this equation should be able to distinguish between the zero value-added and curriculum-matching models. The zero value-added model predicts no discontinuity across different achievement measures. In the curriculum-matching model, Prediction 4 implies upward jumps at the cutoffs, particularly on the math-derived outcomes.

Another feature of this test is that it can potentially difference out peer effects, removing them as a potential alternate explanation for any observed discontinuity. To see why, consider a version of our reduced-form curriculum-matching model where the student's score on test k is affected by match quality, school quality, and peer effects, as follows:

$$y_{ik}(\theta_i, c) = h_k(\theta_i) + m_k(\theta_i, c) + \tau_s + b \sum_{j \neq i} w_{ji}\theta_j + \epsilon_{ik}. \tag{5}$$

Evidence of an upward jump from a standard RD estimate on one test score could be due to any combination of three factors: a positive school effect ($\tau_1 > \tau_2$), positive peer effects ($b > 0$), or students just above the cutoff being better matched to their school's curriculum ($m(\hat{\theta}, c_1) > m(\hat{\theta}, c_2)$). If the dependent variable is the difference between scores on a hard and an easy test, the school effect and peer effect terms will both drop out **if** the school quality and peer effect terms apply equally to the two tests; that is, when τ_s and b do not depend on the test k as we've written in equation (5). This leaves only the effect of how well-matched each school's curriculum is to the hard vs. the easy test for a student of type $\hat{\theta}$ as a potential explanation.[24]

Figure 7 presents graphical estimates for each outcome. BLS applicants who scored within 20 percentiles of the cutoff are grouped into one-percentile bins. The dots give the average within each bin of the difference between students' z-scores on the PSAT/SAT and their z-score on the MCAS. The fitted curves are local linear regression estimates of the conditional mean. The top part of the figure presents estimates from the two math tests. The PSAT figure on the left has a clear and quite large ($\sim 0.2\sigma$) upward jump at the Boston Latin cutoff. The SAT figure on the right has a smaller but still quite visually clear upward jump as well. The bottom half of the figure shows corresponding figures

[23] For each outcome, we compute the z-score among the population of all applicants who rank BLS first within each year who have the test outcomes, and then take the difference between the two z-scores.

[24] Once again, under the null of no curriculum-matching effects, the model implies that the second derivative of the score difference-entrance exam score relationship would be the same to the left and right of the cutoff, mitigating concerns about bias when window widths are not infinitesimal. There is also no reason to think that $\mathbb{E}(h_k(\theta)|r)$ would be highly nonlinear around the cutoff. Students at the cutoff have MCAS scores around 70th percentile, so we are not in the tail of the distribution.

derived from PSAT reading and SAT critical reading scores. We anticipated that these would not be as large given that the MCAS English test does not seem comparably below grade level, and there is little evidence of a discontinuity in the PSAT-based figure. The SAT critical reading based figure suggests an upward jump a bit smaller than those in the math-based figures. Taken together, the visible jumps in three of the four cases suggest that the data are more consistent with the curriculum-matching model than with the zero value-added model (or with a model in which a negative school value-added is offset by a positive peer effect.)

To assess the statistical significance of the jumps in the figure, Table 3 presents RD estimates of the jumps at the cutoff with standard errors obtained from local linear estimates of equation (4) with a score difference as the dependent variable. The PSAT math-MCAS math and the SAT math-MCAS math derived dependent variables jump up by about 0.2σ at the cutoff. The SAT critical reading-MCAS English derived dependent variables jump by about 0.15σ. Each is significant at the 1% level. We conclude that considering the differences between how students perform on more- and less-difficult tests is more consistent with the curriculum-matching model than with the zero value-added model.

To look further performance on more difficult assessment, we investigate BLS's effect on Advanced Placement (AP) test outcomes. The College Board's AP program allows high school students to take college-level courses, and potentially earn college credit based on the AP test performance. AP tests are scored between 1-5, with some colleges granting credit for scores of 3 or higher, which we will refer to as a passing grade. In the last two decades, the number of AP subjects and AP exams administered has increased. Following Abdulkadiroğlu, Angrist, and Pathak (2014), we also consider performance on the most popular AP exams, defined by those taken by at least 500 or at least 1,000 students at Boston Public Schools. The restriction amounts to a focus on core subjects like math, science, English, history, and omits arts, music, and foreign languages. Table 1 shows that Boston 6th graders eventually take an average of 0.5 AP tests, while applicants near the BLS cutoff average 1.8. Moreover, BLS applicants near the admissions cutoff are more likely to take and pass popular AP tests. For example, the fraction of students who take AP exams taken by at least 500 students is four times larger for applicants near the BLS cutoff than for the general Boston student population.

BLS students account for a large share of AP tests taken by all Boston students. In 2022, BLS students took a total of 2,341 AP tests and scored a three or higher on 84.2% of them. This represents 38% of all AP tests taken by Boston Public School students and 58% (1971 of 3379) of the passing scores, even though BLS accounts for about 10% of overall BPS high school enrollment. The concentration is even larger when considering challenging AP exams taken by high school seniors: BLS

accounts for 82% of the passing scores on the AP Calculus BC exam and 84% of the passing scores on the AP English Literature exam.[25] Consistent with our curriculum-matching model, Boston Latin is able to offer these advanced subjects given its high-achieving students.

Given the discreteness of AP achievement, it would not be natural to normalize AP achievement as a z-score. Accordingly, our AP estimates focus on the effect of BLS admissions on AP test-taking and performance, rather than on a differenced dependent variable that subtracts normalized MCAS performance from normalized AP performance.

BLS admission increases AP test taking, test passing rates, and scores for our sample of applicants. The first can be seen in Panel A of Table 4, which shows that a BLS offer results in an increase of 0.9 AP exams taken, and an increase of 0.4 for popular AP tests (measured either with tests with 500+ or 1000+ takers). The magnitude of these effects are large compared to the control mean. Panel B shows that a BLS offer increases the number of AP tests passed by between 0.17 and 0.13, depending on what tests are counted.[26] We also measure AP test performance by an adjusted sum of AP scores, which only counts scores when they are three or higher. Specifically, we adjust AP scores by giving a 5 an adjusted score of 3, a 4 an adjusted score of 2, a 3 an adjusted score of 1, and adjusting all other scores to 0. Panel C shows that a BLS offer increases the adjusted sum of scores on all AP exams by 0.8, and increases it by about 0.3 for the popular tests. The consistent pattern of effects of BLS on AP test-taking and performance support the curriculum-matching model. A selective school can offer an advanced curriculum like Advanced Placement courses when high-ability students are grouped together at a single school. The AP results here are not consistent with the zero-value added model.

The AP analyses also provide evidence of slope-changes at the BLS cutoff consistent with Prediction 1. The regressions with test-taking as the dependent variable have positive significant slope-change coefficient estimates regardless of whether we consider all AP exams or either of the subsets of subjects. The same is true of the regressions examining the adjusted sum of scores. This is consistent with students who are further above the admissions cutoff being better matched to the very rigorous AP courses that BLS offers. Over 80% (74 of 91) of the BLS students who took the AP Calculus BC exam in 2023 received the highest possible score of 5. This is extremely impressive and indicates that

[25] These statistics are available at: https://profiles.doe.mass.edu/adv_placement/ap.aspx

[26] Abdulkadiroğlu, Angrist, and Pathak (2014) report estimates of Boston exam school attendance on AP test-taking and AP scores for 7th and 9th grade applicants. Those estimates much noisier than those reported here. For example, the estimated standard error on sum of scores is 0.61 and 0.48 for AP tests with 500+ takes compared to standard errors of 0.25 and 0.18 in Table 4. The implied confidence intervals for this outcome in Abdulkadiroğlu, Angrist, and Pathak (2014) are wide enough to include the estimates reported here.

the course was taught at a very high level, but also suggests that additional BLS students might have enrolled in Calculus BC and earned a passing grade if the school's course was a bit less demanding.

5 A Postscript on College

As noted earlier, BLS views itself as a college-preparatory school. Low rates of college completion is an enormous issue for Boson students (as it is for students nationwide). Even restricting an analysis to the relatively high-achieving students with admissions test score in the window around the BLS cutoff, Table 1 indicates that over 50% of those who start college fail to graduate. Given its college-preparatory mission and low graduation rates, BLS might well regard its students' success in college as a more apt measure of the effect it hopes to have than are scores on SAT, AP, or other tests.

BLS does appear to increase four-year college graduation, although the result is only marginally significant. The point estimates on the probability of graduation in Table 5 are that there are jumps of 6.3 percentage points in graduation from any college and 7.5 percentage points in graduation from a 4-year college at the BLS cutoff. The latter is significant at the 10% level. An effect of the size of the point estimates would be quite large practically given that the control mean probability of graduating from a 4-year college is just 34%.

The probability of college graduation depends both on college attendance and on the probability of graduation conditional on attendance. BLS is a feeder school for many elite colleges. For example, Bernhard (2013) reports that BLS sent 13 students to Harvard in 2013, placing it among the nation's top seven feeder high schools. Since selective colleges often limit the number of offers made to applicants from any given high school, however, it is unclear one would expect BLS to have an effect on the likelihood of admission to elite colleges for students who were in the left tail when admitted to BLS. Abdulkadiroğlu, Angrist, and Pathak (2014) reported a positive and marginally significant 6.2 percentage point effect of BLS admissions on college-going, and a 10 percentage point boost on four-year college attendance, significant at the 5% level. In our sample, which has several additional years but restricts to students applying as 6th graders, the estimated admissions effects are much smaller and not statistically significant, although we also cannot reject that there are jumps of the size they estimated. We do get a significant slope-change estimate, suggesting a positive effect on 4-year college enrollment for students further above the BLS cutoff.

6 Conclusion

This paper develops two simple models of optimal school system and curriculum design. A school system serving students with different learning styles maximizes aggregate achievement by assigning students to different schools based on their learning style. In the model, schools allocate time to teaching different skills based on the learning styles of their students. The model implies that RD estimates of the effect of selective schools on a test score outcome do not provide a way to adjudicate whether student allocation/curriculum choices are well-designed. Our framework shows that no effect at an admissions cutoff is consistent with two very different models of educational production: (1) the more selective school has zero value-added; or (2) the student allocation and the curriculum design are both chosen optimally to maximize aggregate achievement.

Our message is not a negative one, however. The curriculum-matching model generates predictions that can differentiate between alternative models. Some key differences are that the optimal curriculum-matching predicts (1) changes in the slope of the entrance-exam later performance relationship across admissions cutoffs (because students are shifted from being worse to better matched to their schools' curriculum); (2) students should express an ex-ante preference for the more selective school (because the preference only matters if they score well enough to be admitted); (3) discontinuities in individual question scores should exist (because schools allocate different amounts of time to different topics); and (4) students at the more selective school will do better on more challenging tests in a neighborhood of admissions cutoffs (because more selective schools focus on skills assessed by more challenging tests).

Investigating some of the predictions of the curriculum-matching model presents challenges. For example, the estimation of a slope change at an admissions cutoff requires more data than estimating a change in levels. Moreover, measuring performance on individual test questions is inherently noisy because the outcomes entail a 0-1 dependent variable and cover just one year of data since test questions differ each year. Nonetheless, we find support for almost all of the predictions of the curriculum-matching model. The strongest support comes from what we would ex ante would have expected to be one of the most powerful tests. Students who clear the Boston Latin School admissions cutoff score almost 0.2 standard deviations higher when we use the difference between SAT performance and 10th grade MCAS performance as our measure of achievement.

We developed our curriculum-matching model focusing on a comparison of schools that placed the same value on all skills and specialized only because different students learned the skills at different

rates. In practice, it may be equally important that the value of mastering a skill varies across students. College-bound students may benefit from being exposed to instructional material with a complexity level of college texts or having the math background needed for college-level STEM courses. For other students, it may be more important to master topics covered on graduation exit exams that may be directly relevant to everyday life.

Another lesson of our analysis is that studies that use test scores to evaluate educational interventions should carefully consider the alignment between the measures they are using and the aims of the educational intervention. Tests can differ both in terms of tested content and in how mastery of the material translates to a numeric score. Schools teach many different skills that provide short and long-term benefits to students. Our analysis shows how ignoring this consideration may lead to a misleading assessment of the value of selective schools and provide an important caution for evaluations of other educational interventions.

References

Abdulkadiroğlu, A., J. Angrist, and P. Pathak (2014): "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," *Econometrica*, 82(1), 137–196.

Abdulkadiroğlu, A., J. D. Angrist, Y. Narita, and P. A. Pathak (2022): "Breaking Ties: Regression Discontinuity Design Meets Market Design," *Econometrica*, 90(1), 117–151.

Barry, E. (2021): "Boston Overhauls Admissions to Exclusive Exam Schools," NY Times, July 15.

Bernhard, M. P. (2013): "The Making of a Harvard Feeder School," The Harvard Crimson, December 13, Available at: https://www.thecrimson.com/article/2013/12/13/making-harvard-feeder-schools/.

Bleemer, Z. (2024): "Top Percent Policies and the Return to Postsecondary Selectivity," Working Paper, Princeton University.

BLS (2016): "Boston Latin School Student Handbook, 2016-2017," Available at: https://www.bls.org/downloads/StudentInfo/BLS%20HANDBOOK%2016-17.pdf.

Bui, S., S. Craig, and S. Imberman (2014): "Is Gifted Education a Bright Idea? Assessing the Impacts of Gifted and Talented Program," *American Economic Journal - Economic Policy*, 6(3).

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

Card, D., and L. Guiliano (2016): "Can Tracking Raise the Test Scores of High-Ability Minority Students?," *American Economic Review*, 106(1), 2783–2816.

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on CAusal Effects in a Generalized Regression Kink Design," *Econometrica*, 83(6), 2453–2483.

Chetty, R., D. J. Deming, and J. N. Friedman (2023): "Diversifying Society's Leaders? The Determinants and Causal effects of Admissions to Highly Selective Private Colleges," NBER Working paper, 31492.

Cohodes, S. (2020): "The Long-Run Impacts of Tracking High-Achieving Students: Evidence from Boston's Advanced Work Class," *American Economic Journal: Economic Policy*, 12(1).

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

Dale, S., and A. B. Krueger (2002): "Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables," *Quarterly Journal of Economics*, 117(4), 1491–1527.

Deming, D. (2023): "Multidimensional Human Capital and the Wage Structure," Chapter prepared for the Handbook of the Economics of Education.

Dobbie, W., and R. G. Fryer (2014): "Exam High Schools and Academic Achievement: Evidence from New York City," *American Economic Journal: Applied Economics*, 6(3), 58–75.

Duflo, E., P. Dupas, and M. Kremer (2011): "Peer Effects and the Impacts of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101(5), 1739–1774.

Ellison, G. (2010): *Hard Math for Middle School*. CreateSpace Independent Publishing Platform.

——— (2013): *Hard Math for Elementary School*. CreateSpace Independent Publishing Platform.

Ganong, P., and S. Jäger (2018): "A Permutation Test for the Regression Kink Design," *Journal of the American Statistical Association*, 113:522, 494–504.

Hiebert, E. H. (2009): "Interpreting Lexiles in Online Contexts and with Informational Texts," Apex Learning.

Hiebert, E. H., and H. A. E. Mesmer (2013): "Upping the Ante of Text Complexity in the Common Core State Standards: Examining Its Potential Impact on Young Readers," *Educational Researcher*, 42(1), 44–51.

Hindley, B., and V. Eaton (2018): "Staying the Course: Six-Year College Enrollment and Completion of BPS Class of 2011 Graduates," The Boston Private Industry Council.

Hoekstra, M. (2009): "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach," *Review of Economics and Statistics*, 91(4), 717–724.

Imbens, G., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79(3), 933–959.

Jackson, K. (2018): "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes," *Journal of Political Economy*, 126(5), 2072–2107.

Jacob, B. A. (2005): "Accountability, incentives and behavior: the impact of high-stakes testing in Chicago Public Schools," *Journal of Public Economics*, 89, 761–796.

Lazear, E. (2001): "Educational Production," *Quarterly Journal of Economics*, 116(3), 777–803.

Mountjoy, J., and B. R. Hickman (2021): "The Returns to College(s): Relative Value-Added and Match Effects in Higher Education," BFI Working Paper 29276.

Pop-Eleches, C., and M. Urquiola (2013): "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, 103(4), 1289–1324.

Urquiola, M., and E. Verhoogen (2009): "Class-Size Caps, Sorting and the Regression-Discontinuity Design," *American Economic Review*, 99(1), 179–215.

Zimmerman, S. (2014): "The Returns to College Admission for Academically Marginal Students," *Journal of Labor Economics*, 32(4), 711–754.

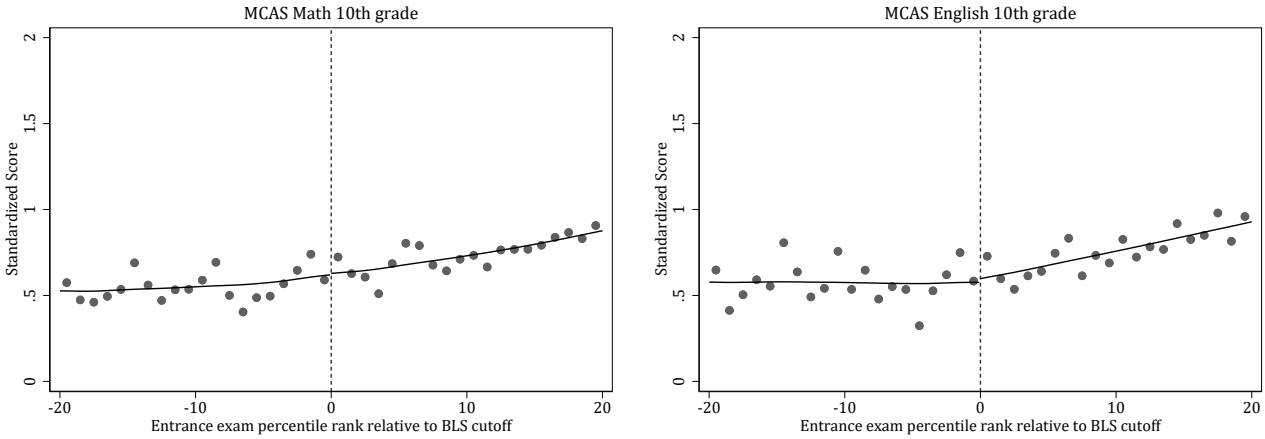


Figure 1: Performance on 10th grade MCAS vs. admissions test score.

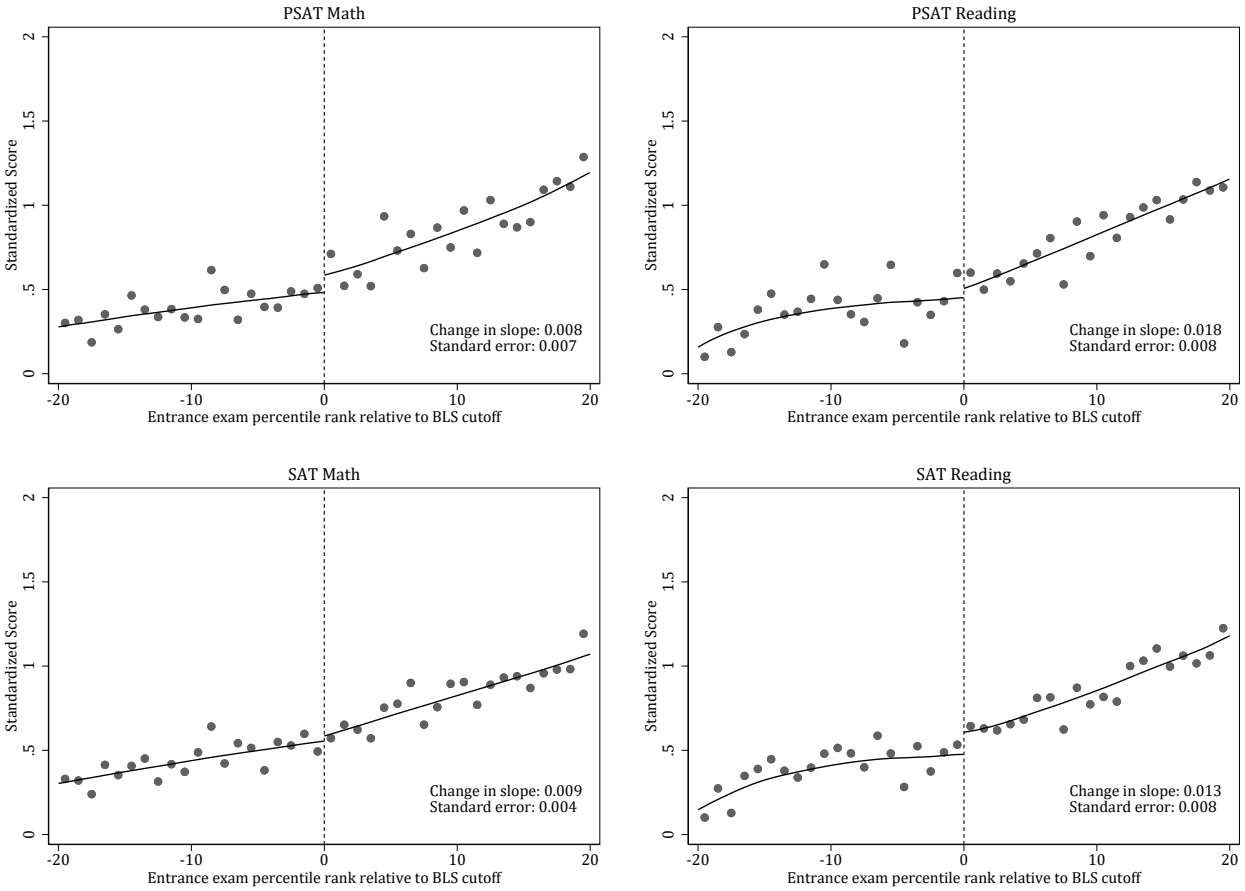


Figure 2: Slopes of the achievement vs. entrance score relationship around the admissions cutoff

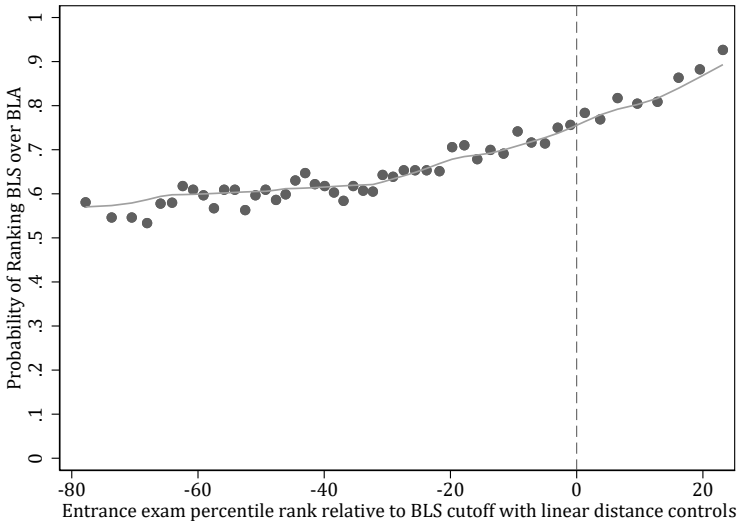


Figure 3: Fraction of students preferring Boston Latin School (BLS) to Boston Latin Academy (BLA) by admissions exam percentile relative to the BLS cutoff

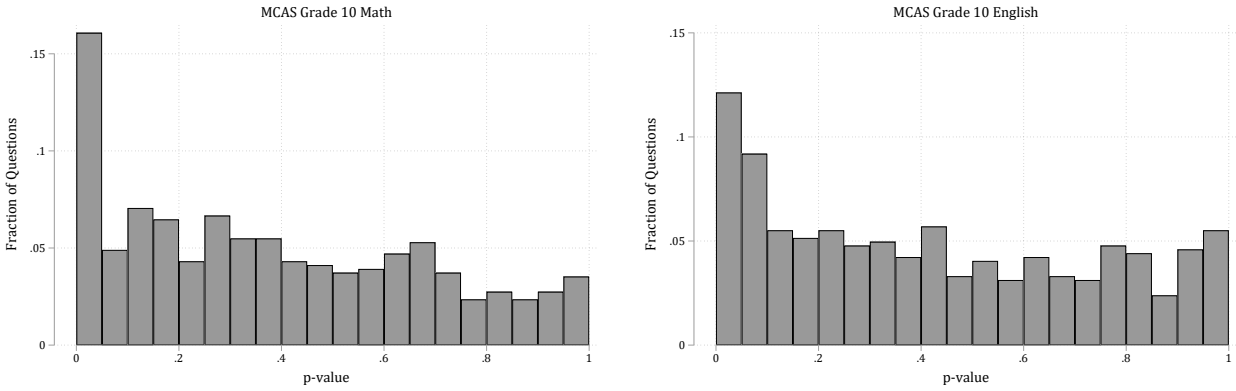
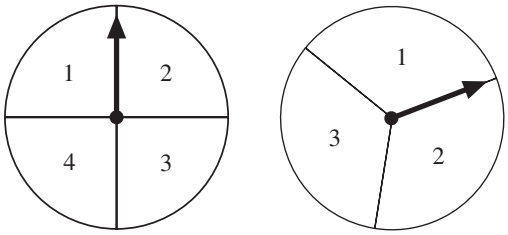


Figure 4: P-values of jumps at the Boston Latin School cutoff on individual grade 10 MCAS questions.

10 Joshua spun the arrow on each spinner shown below exactly once. He recorded the sum of the resulting two numbers.



What is the probability that the sum of the resulting two numbers will be 2?

28 Two containers in the shape of right circular cylinders are equal in height. The radius of the larger container is 3 times the radius of the smaller container.

The volume of the larger container is how many times the volume of the smaller container?

30 The length of a rectangle is 1 inch more than 2 times its width. The area of the rectangle is 36 square inches.

What is the **length** of the rectangle?

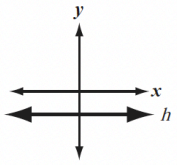
A. 4 inches
B. 6 inches
C. 9 inches
D. 18 inches

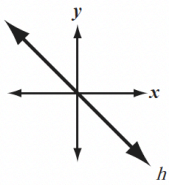
1 The first five terms in a quadratic sequence are shown below.

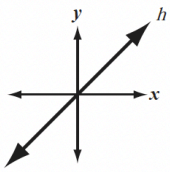
6, 9, 14, 21, 30, . . .

What is the next term in the sequence?

34 In which of the following graphs does line h best represent a line with an undefined slope?

A.


B.


C.


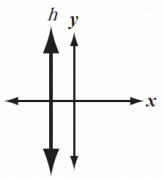
D.


Figure 5: Sample MCAS questions with downward jumps at the Boston Latin School cutoff

11 If $y \neq 0$, which of the following is equivalent to the expression below?

$$\frac{15y^9}{5y^3}$$

13 Which of the following is equivalent to the expression below?

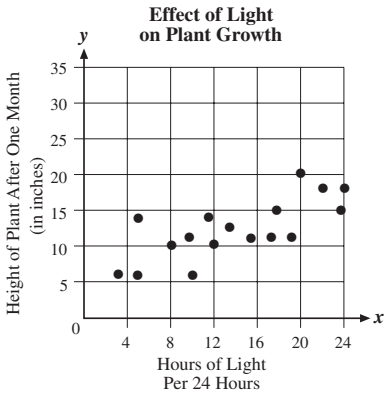
$$-2(x - 5)$$

5 Ms. Burke correctly weighed a tomato to the nearest ounce and recorded the weight. The weight she recorded was 13 ounces.

What is the **least** possible actual weight of the tomato?

A. 12.0 ounces
B. 12.5 ounces
C. 13.0 ounces
D. 13.4 ounces

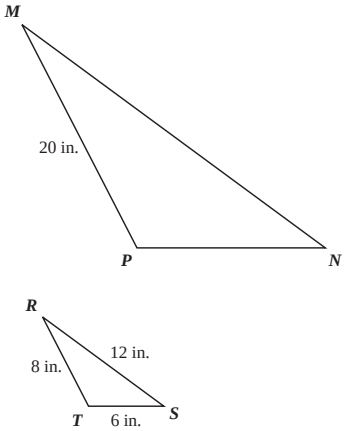
39 Jenny studied the effect of light on plant growth. She graphed a scatterplot to represent her data.



Which of the following **best** represents the equation for the line of best fit for the data shown?

A. $y = -0.4x + 5$
B. $y = 0.4x + 5$
C. $y = -4x + 5$
D. $y = 4x + 5$

30 In the diagram below, $\triangle MNP \sim \triangle RST$.



Based on the dimensions in the diagram, what is the length of $\overline{MN}$?

Figure 6: Sample MCAS questions with upward jumps at the Boston Latin School cutoff

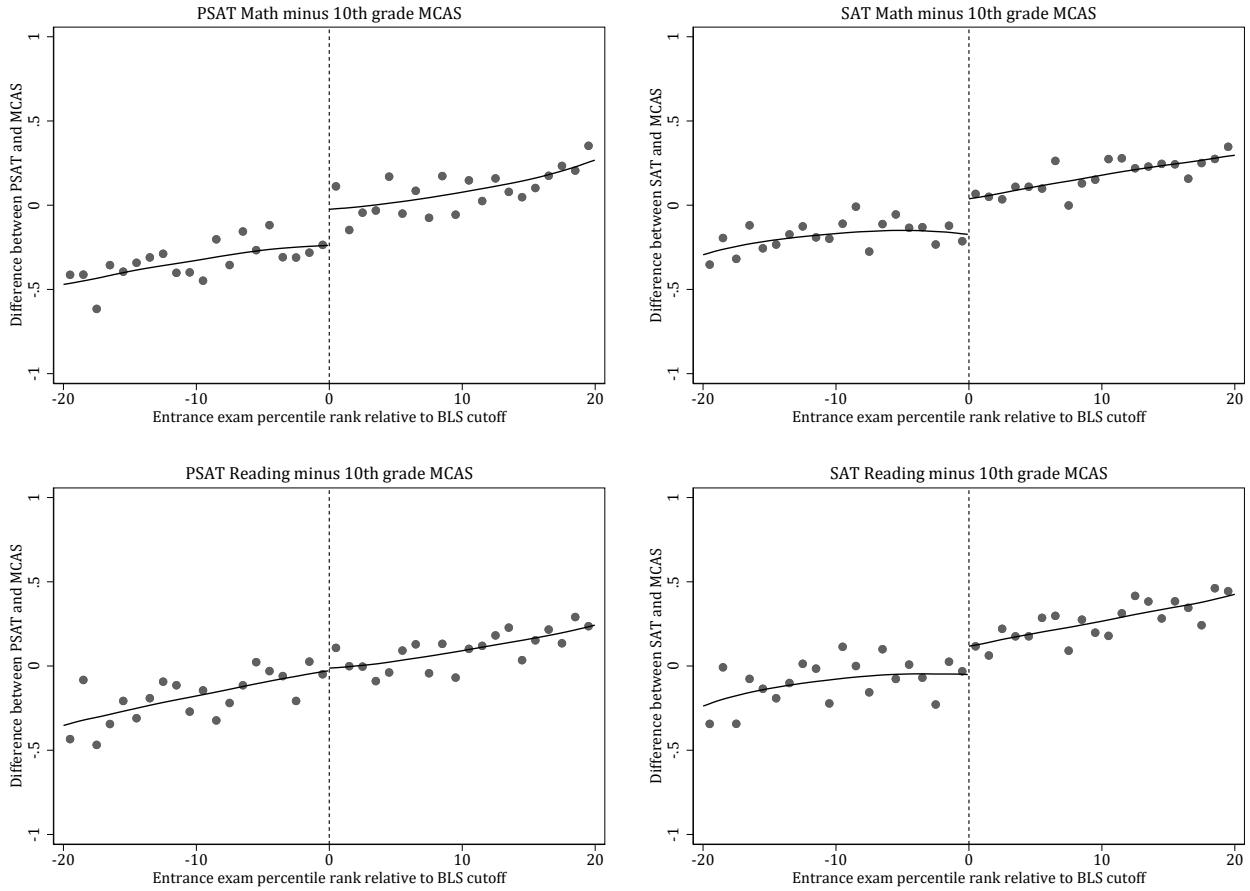


Figure 7: Differences between scores on more and less difficult tests: PSAT/SAT score minus 10th grade MCAS score at the Boston Latin School cutoff

Table 1: Descriptive Statistics for 6th Grade BPS Students and BLS Applicants

	BPS 6th graders	6th grade BLS Applicants	6th grade Applicants Ranking BLS First	6th grade BLS Applicants in [-20,20] window
	(1)	(2)	(3)	(4)
		A. Demographics		
Black	0.45	0.38	0.32	0.22
Hispanic	0.33	0.28	0.26	0.22
White	0.13	0.16	0.19	0.25
Asian	0.08	0.17	0.21	0.31
		B. MCAS test outcomes		
6th grade Math	-0.48	0.01	0.16	0.70
6th grade English	-0.55	-0.13	0.01	0.38
10th grade Math	-0.36	0.18	0.32	0.76
10th grade English	-0.47	-0.01	0.12	0.49
		C. Pre-college test outcomes		
PSAT Math	41	44	46	50
PSAT Verbal	38	40	42	46
SAT Math	471	511	532	578
SAT Verbal	447	482	503	544
Number of AP Taken	0.54	1.05	1.30	1.77
AP Score ≥ 3	0.09	0.20	0.26	0.40
AP 500+ Score ≥ 3	0.07	0.16	0.21	0.32
AP 1000+ Score ≥ 3	0.06	0.13	0.18	0.27
		D. Post-secondary outcomes		
Attend Any College	0.56	0.75	0.77	0.85
Attend 4 yr College	0.38	0.60	0.64	0.77
Persist Any College	0.48	0.67	0.69	0.77
Persist 4 yr College	0.34	0.54	0.58	0.70
Graduate Any College	0.16	0.27	0.30	0.42
Graduate 4 yr College	0.13	0.24	0.28	0.40

Notes: This table presents characteristics of Boston 6th graders and samples of BLS applicants. Applicants in [-20,20] window have running variable values within twenty percentile ranks of the admissions cutoff. AP tests with 1000+ takers include U.S. History, Biology, English Language and Composition, English Literature and Composition, U.S. Government and Politics, and Calculus AB. AP tests with at least 500 takers include all the tests with at least 1,000 takers, as well as Calculus BC, Statistics, Chemistry, Physics B/C, European History, Microeconomics, and Macroeconomics. AP Score ≥ 3 counts the number of AP tests with a 3 or higher. AP 500+ Score ≥ 3 counts the number of AP tests with 500+ takers with a 3 or higher, and similarly for AP 1000+ Score ≥ 3. Any college refers to either a 2 or 4-year college. Persist means attend any college for at least four semesters.

Table 2: Discontinuities in Achievement-Entrance Score Relationship at Admissions Cutoffs

	Math			English		
	MCAS	PSAT	SAT	MCAS	PSAT	SAT
	(1)	(2)	(3)	(4)	(5)	(6)
Change in level at cutoff ($\boldsymbol{\rho}$)	0.021	0.118*	0.074*	0.003	0.076	0.185***
	(0.047)	(0.064)	(0.044)	(0.050)	(0.066)	(0.056)
Change in slope at cutoff ($\boldsymbol{\gamma}$)	-0.005	0.008	0.009**	0.007**	0.018**	0.013*
	(0.005)	(0.007)	(0.004)	(0.004)	(0.008)	(0.008)
Control mean	0.43	0.42	0.38	0.49	0.4	0.39
Observations	3028	2173	3959	4592	1939	2230
Bandwidth	18.11	17.73	23.68	27.65	15.9	13.61

Notes: This table reports estimates of changes in levels and the slope of achievement outcomes at the BLS admissions cutoff. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table also includes the number of observations, the bandwidth, and the control mean scores for comparison.

Table 3: Discontinuities in SAT/PSAT Relative to MCAS at Admissions Cutoffs

	Difference in Math		Difference in English	
	PSAT-MCAS	SAT-MCAS	PSAT-MCAS	SAT-MCAS
	(1)	(2)	(3)	(4)
Change in level at cutoff ($\boldsymbol{\rho}$)	0.217***	0.198***	0.022	0.147**
	(0.066)	(0.056)	(0.075)	(0.067)
Control mean	0.000	0.020	0.060	0.110
Observations	1587	2063	1975	2183
Bandwidth	16.26	14.97	20.13	15.91

Notes: This table displays the changes in differences in academic achievement measured as the difference between a student's score on the math/reading PSAT and the 10th grade math/English MCAS and the difference between a student's score on the math/critical reading SAT and the 10th grade math/English MCAS. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Advanced Placement Test Outcomes at Admissions Cutoffs

	All AP Exams	AP Exams with 500+ Takers	AP Exams with 1000+ Takers
	(1)	(2)	(3)
		A. Test Taking	
Change in level at cutoff (ρ)	0.885***	0.402***	0.402***
	(0.132)	(0.109)	(0.084)
Change in slope at cutoff (γ)	0.053***	0.034***	0.029**
	(0.013)	(0.012)	(0.012)
Control mean	1.75	1.25	0.8
Observations	3453	3221	2506
Bandwidth	20.49	19.15	14.86
		B. Number of AP Tests Passed	
Change in level at cutoff (ρ)	0.170***	0.127***	0.148***
	(0.032)	(0.040)	(0.039)
Change in slope at cutoff (γ)	0.008***	0.006	0.009
	(0.003)	(0.006)	(0.006)
Control mean	0.35	0.29	0.25
Observations	3912	2434	2406
Bandwidth	23.19	14.45	14.28
		C. Adjusted Sum of Scores	
Change in level at cutoff (ρ)	0.817***	0.315*	0.376***
	(0.245)	(0.177)	(0.115)
Change in slope at cutoff (γ)	0.116***	0.079***	0.061***
	(0.042)	(0.030)	(0.021)
Control mean	1.56	1.02	0.62
Observations	2076	2123	1943
Bandwidth	12.29	12.61	11.45

Notes: Panel A shows the change in the level and slope of the number of students taking AP exams at the cutoff for all AP tests, and AP tests with 500+ and 1000+ takers. Tests with at least 1,000 takers include U.S. History, Biology, English Language and Composition, English Literature and Composition, U.S. Government and Politics, and Calculus AB. Tests with at least 500 takers include all the tests with at least 1,000 takers, as well as Calculus BC, Statistics, Chemistry, Physics B/C, European History, Microeconomics, and Macroeconomics. Panel B reports the change in the number of students passing AP exams defined as scoring 3 or higher at the cutoff. Panel C presents change in the adjusted sum of AP scores at the admissions cutoff, where scores are adjusted to assign 3 points for a score of 5, 2 points for a score of 4, and 1 point for a score of 3. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: College Outcomes at Admissions Cutoffs

	Attend	Persist	Graduate
	(1)	(2)	(3)
		A. Any College	
Change in level at cutoff (ρ)	0.003	0.019	0.063
	(0.030)	(0.031)	(0.041)
Change in slope at cutoff (γ)	0.005	0.003	0.006
	(0.005)	(0.005)	(0.005)
Control mean	0.81	0.71	0.36
Observations	6031	6156	6323
Bandwidth	13.1	13.38	16.68
		B. 4 Year College	
Change in level at cutoff (ρ)	0.034	0.01	0.075*
	(0.039)	(0.042)	(0.041)
Change in slope at cutoff (γ)	0.016**	0.016**	0.008
	(0.006)	(0.008)	(0.005)
Control mean	0.70	0.61	0.34
Observations	5611	4982	6187
Bandwidth	12.16	10.74	16.29

Notes: This table reports the change in the level and slope of the number on college attendance, persistence, and graduation rate. Persist means attend any college for at least four semesters. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A Appendix

A.1 Proof of Proposition 7

Since the functions $t_1^*(x)$ and $t_2^*(x)$ are continuous in x and have the same integral over $[0,1]$ there must be an $x_0 \in [0,1]$ with $t_1^*(x_0) = t_2^*(x_0)$.

Given any x with $xv(x) < x_0 v(x_0)$ define $\hat{t}(x)$ to be the allocation of time to skill x that would equate the marginal value of time spent studying x and x_0 for a student of type $\hat{\theta}$ (provided such a nonzero amount exists), i.e.

$$\hat{\theta} x v(x) e^{-\hat{\theta} x \hat{t}(x)} = \hat{\theta} x_0 v(x_0) e^{-\hat{\theta} x_0 t_2^*(x_0)}.$$

We show below that for any $\theta < \hat{\theta}$,

$$x v(x) e^{-\theta x \hat{t}(x)} < x_0 v(x_0) e^{-\theta x_0 t_2^*(x_0)}.$$

Integrating over the interval $[0, \hat{\theta}]$ gives

$$\int \theta x v(x) e^{-\hat{\theta} x \hat{t}(x)} g_2(\theta) d\theta < \int \theta x_0 v(x_0) e^{-\theta x_0 t_2^*(x_0)} g_2(\theta) d\theta.$$

This equation implies that the marginal value of time spent teaching skill x at school 2 would be less than the marginal value of time spent teaching skill x_0 at school 2 if the times spent on those skills were $\hat{t}(x)$ and $t_2^*(x_0)$, respectively. Hence, the optimal amount of time spent on skill x at school 2 must satisfy $t_2^*(x) < \hat{t}(x)$.

A similar argument implies that $t_1^*(x) > \hat{t}(x)$. In combination these two imply that $t_1^*(x) > t_2^*(x)$ which finishes the proof for skills x with $xv(x) < x_0 v(x_0)$. The argument for the $xv(x) > x_0 v(x_0)$ is analogous.

To complete the proof, we need only show the inequality whose proof we deferred:

$$x v(x) e^{-\theta x \hat{t}(x)} < x_0 v(x_0) e^{-\theta x_0 t_2^*(x_0)}$$

for $\theta < \hat{\theta}$. To see this note that

$$\frac{e^{-\theta x \hat{t}(x)}}{e^{-\theta x_0 t_2^*(x_0)}} = \left(\frac{e^{-\hat{\theta} x \hat{t}(x)}}{e^{-\hat{\theta} x_0 t_2^*(x_0)}} \right)^{\theta/\hat{\theta}} = \left(\frac{\hat{\theta} x_0 v(x_0)}{\hat{\theta} x v(x)} \right)^{\theta/\hat{\theta}} < \frac{x_0 v(x_0)}{x v(x)},$$

with the final step coming from the combination of $\theta/\hat{\theta} < 1$ and $\frac{x_0 v(x_0)}{x v(x)} > 1$. Multiplying through by the denominators gives the desired inequality. ■

A.2 MCAS difficulty and grade-level alignment

While Massachusetts has among the highest test scores and standards in the nation, MCAS primarily tests material that students learn in earlier grades. The clearest evidence in support of this claim comes from data from the Massachusetts Department of Elementary and Secondary Education (DESE). DESE provided question-by-question mappings of the 10th grade Math MCAS exam to the common-core aligned 2011 Massachusetts curriculum framework. This maps the majority of 10th grade questions to middle school standards. On the 2013 10th grade Math exam, for example, DESE linked 23 of 42 questions to middle school standards. For example, question 3, which asked "What is the value of $\frac{1}{3} \cdot 6\left(4+9+\sqrt{4 \cdot 9}\right)$?" is mapped to a 7th-grade standard, *7.EE.3 Solve multi-step real-life and mathematical problems posed with positive and negative rational numbers in any form (whole numbers, fractions, decimals), using tools strategically.* Question 10, which provided a 10-observation histogram showing areas in square miles and asked "What is the median area, in square miles, of the towns in the county?", is mapped to a 6th-grade standard, *6.SP.5 Summarize numerical data sets in relation to their context, such as by ... Giving quantitative measures of the center (median and/or mean).*

Comparable MCAS math reports from some earlier years, e.g. 2003, mapped almost all questions to 10th grade standards. However, we do not think that this correspondence reflects the fact that the grade-level of the material covered on the MCAS has varied. Instead, it appears that in those years, an earlier-grade standard was listed when no 10th-grade standard included the topic. To quantify this, we went through the 2003 10th grade Math exam and compared the alignment of each question with then-prevailing (2001) Massachusetts curriculum standards for all grades. For example, we noted that question 1, "A landing pad for a helicopter is in the shape of a circle with a radius of 7 meters. Which of the following is closest to the area of the landing pad?" could be mapped to the 6th grade standard *6.M.5 Identify, measure, and describe circles and the relationships of the diameter, circumference, and area ... and use the concepts to solve problems*, whereas the 2003 MCAS document mapped it to *10.M.1 Calculate perimeter, circumference, and area of common geometric figures such as parallelograms, trapezoids, circles, and triangles.* Across the full 2003 test, we found 25 of 42 questions on the 10th grade tests could be mapped to elementary or middle school standards, a similar rate to DESE reports in 2013.[27]

While it is harder to quantify "difficulty" than "grade level," we believe that it is also true that MCAS questions are fairly easy questions (easier than SAT questions) on the content they are covering. To illustrate this, we reproduce in Figure A1 the **three most difficult** of the 36 short answer/multiple choice questions from the 2013 test, based on the fraction who responded correctly. The first, question 35, requires that students factor quadratic equations to simplify a fraction, but note that the quadratics have been chosen to be about as easy to factor as possible and it is also easy to see, e.g. by plugging in $x = 1$, that three of the four choices could not possibly be correct. The second, question 37, asks students to find the relationship between distance and time. This question could also be solved quickly and easily without using algebra. The question tells you that the formula is supposed to give $1\frac{2}{3}$ as

[27] We mapped one question to a 4th grade standard, two to 5th grade standards, seven to 6th grade standards, six to 7th grade standards, nine to 8th grade standards, and seventeen to 10th grade standards. (Massachusetts had not at the time published corresponding 9th grade standards.)

the answer when you plug in $t = \frac{2}{3}$. The answer choices were selected so that if you plug in $\frac{2}{3}$ of an hour for t, only one formula gives $1\frac{2}{3}$ miles for d. The final question, question 16, does require that students remember order-of-operations from 7th grade, but the calculations, $60 \div 4 = 15$, $15 \times 3 = 45$, and $100 - 45 = 55$, are designed to be easy computations.

Students are not under any time pressure when attempting these questions: the test is untimed and students can spend as long as they like on each section. Most other questions on the test were much easier than these. The average score statewide on these questions was only 40% correct, whereas 71% of students got the correct answer on the median-difficulty question. The PSAT and SAT are more speed-oriented, e.g. the PSAT requires that students complete 44 questions in 70 minutes, and also include some easy questions, e.g. the PSAT's official practice exam includes "What is 23% of 100?" But, the peak difficulty is clearly above those of the MCAS examples above. For example, questions on the PSAT's released practice exam include asking for the positive solution to $5x^2 - 27x + 24$, and asking for the value of k for which the sum of the solutions of the equation $64x^2 - (16a + 4b)x + ab = 0$ is $k(4a + b)$.

35 Which of the following is equivalent to the expression below for all positive values of x?

$$\frac{x^2 + x - 6}{x^2 + 5x + 6}$$

A. -1

B. $-\frac{6}{11}$

C. $\frac{x - 2}{x + 2}$

D. $\frac{(x + 3)(x - 2)}{(x + 6)(x - 1)}$

37 The distance traveled by a student walking at a constant rate varies directly with the amount of time the student walks. The student walked $1\frac{2}{3}$ miles in $\frac{2}{3}$ hour.

Which of the following represents the relationship between d, the distance in miles walked by the student, and t, the amount of time in hours the student walked?

A. $d = \frac{5}{2}t$

B. $d = \frac{7}{3}t$

C. $d = \frac{10}{9}t$

D. $d = \frac{2}{5}t$

16 What is the value of the expression below?

$$100 - 60 \div 4 \cdot 3$$

Figure A1: The three most difficult questions on the 2013 MCAS grade 10 math test

Quantifying the grade level of material covered on English tests is more difficult. One measure that is available for many classic and school-marketed books due to agreements between its seller, Metametrics, and a number of publishers is the "lexile" measure, which reflects how common the

words in a text are and the average lengths of the sentences.[28] Table A1 list the Lexile indexes we could find for books that had excerpts used as reading passages on 2004-2013 10th grade MCAS English exams. While there is very limited scientific justification for the measure, the 2010 Common Core Standards for the English Language Arts did include a recommendation that the grade 9-10 students read texts in the 1080-1305 range to be on a path to be reading college-level texts in 12th grade.[29] In addition to being skeptical of the measures themselves, we think this seems a little high given that Lexile's promoters at the time recommended texts in the 960 to 1115 range for grades 9-10. In any case, we encourage readers to use the ordered list and their own knowledge of some of the texts to get a sense of the grade level. We feel that the books are close to grade level and certainly much closer than the grade 10 MCAS math problems.

Table A1: Lexile ratings of books excepted on MCAS English tests: 2004-2013

Year	Author	Title	Lexile
2007	Chevalier	Girl with the Pearl Earring	770
2011	Rand	The Fountainhead	790
2011	O'Brien	The Things They Carried	880
2008	Wright	Black Boy	950
2010	Conrad	Heart of Darkness	970
2009	Dickens	Oliver Twist	1000
2013	Zoya (with Follain and Christoforo)	Zoya's Story	1000
2008	Shelley	Frankenstein	1000
2007	Hamilton	Mythology	1040
2006	Lahiri	Interpreter of Maladies	1050
2004	Dickens	Hard Times	1060
2011	Almond	Candyfreak: A Journey through the ...	1080
2012	Unknown (trans. Heaney)	Beowulf	1090
2013	Kafka	The Trial	1150
2005	Austin	Pride and Prejudice	1190
2012	Sullivan	Rats: Observations on the History & ...	1230
2004	Allende (trans. Bogin)	The House of the Spirits	1280
2006	Cervantes (trans. Grossman)	Don Quixote	1410
2008	Ellis	Founding Brothers: The Revolutionary ...	1410
2010	Garcia Marquez (trans. Grossman)	Love in the Time of Cholera	1440

[28] See Hiebert (2009).
[29] See Hiebert and Mesmer (2013).

A.3 Attrition

Table A1: Differential Attrition

	Have MCAS (1)	Have PSAT (2)	Have SAT (3)	Have PSAT-MCAS (4)	Have SAT-MCAS (5)	NSC Queried (6)
				A. Math		
Change in level at cutoff (ρ	0.110***	-0.025	0.028	-0.015	0.050*	
	(0.023)	(0.017)	(0.020)	(0.022)	(0.029)	
Control mean	0.66	0.47	0.68	0.35	0.47	
Observations	7152	7467	9356	3227	2472	
Bandwidth	13.96	14.68	18.34	19.17	14.7	
				B. English		
Change in level at cutoff (ρ	0.106***	-0.025	0.028	-0.022	0.047*	
	(0.019)	(0.017)	(0.020)	(0.021)	(0.027)	
Control mean	0.69	0.48	0.68	0.35	0.47	
Observations	8343	7467	9356	3333	2602	
Bandwidth	16.31	14.68	18.34	19.8	14.7	
				C. Other Outcomes		
Change in level at cutoff (ρ)						-0.001
						(0.006)
Control mean						0.85
Observations						6846
Bandwidth						13.38

Notes: This table reports on outcome availability at admissions cutoffs. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Data Appendix

Lists of exam applicants, their rank order, and school assignments are created using annual records from the Boston Public Schools (BPS) school assignment system. The Massachusetts Student Information Management System (SIMS) is the source for information on student demographic characteristics and school attendance. Achievement test scores come from the Massachusetts Comprehensive Assessment System (MCAS), while college attendance information is provided by the National Student Clearinghouse (NSC). Data on school characteristics were obtained from publicly accessible records from the Massachusetts Department of Elementary and Secondary Education (DESE). SIMS and NSC data were supplied by DESE. SAT, PSAT, and AP exam scores were sourced from both BPS and DESE. Further details on these outcomes are provided below.

BPS Exam School Applicant Data

We received exam school applicant data from BPS for all applicants. The exam school application file contains a record for each applicant, including an application ID number, state ID (SASID) number, name, gender, race, date of birth, application year, grade of application, preferences for the three exam schools, and the composite score for admission. Each record also indicates the school offered to the applicant, if any. The analysis sample includes exam school applicants applying to 7th grade for school years 1999 through 2015. These years correspond to those for which we have available outcome test scores. Application data is supplemented with SIMS enrollment records, which provide geocode data for the application years. Application, enrollment files, test scores, and college outcomes are linked by the SASID applicant identifier.

SIMS Data

Our analysis uses SIMS files for the school years beginning fall 2002-15. These files include information on student demographics, behavioral outcomes, and schools attended for all students enrolled in Massachusetts public schools each year. SIMS contain a state student identifier known as the SASID, used to match SIMS records to MCAS and NSC files.

We use SIMS data to code school enrollment, special education status, subsidized lunch status, and limited English proficiency. Students who attend multiple schools in the same school year are recorded as attending the school they attended longest. Students are classified as qualifying for special education services or eligible for a free or reduced-price lunch in any record within a school-year-grade retain that designation for the entire year-grade.

MCAS Data

MCAS is a state exam used as part of Massachusetts accountability requirements. It is administered in grades 3-10. Our analysis uses MCAS Math and English Language Arts (ELA) data from the school years 2002-2003 to 2019-2020, with scores linked to student SIMS records using SASIDs. Specifically, we focus on 10th grade MCAS ELA and Math test scores. Scaled scores are standardized to have

a mean of zero and a standard deviation of one within each subject-grade-year in our sample who are tested in a given year. A handful of students have a raw score, but no scaled score; for impute the scaled score using the raw-scaled score relationship from other students on the same test. Our analysis also incorporates question-level MCAS data from DESE for every multiple-choice question on the Grade 10 MCAS exam in both Math and ELA for each year, excluding 2016, as question-level data were not available for that year. We count how many students answered each question correctly. It is important to note that the grade 10 MCAS exams were not affected by the Partnership for Assessment of Readiness for College and Careers (PARCC) changes in 2015 and 2016.

SAT Data

The Scholastic Aptitude Test (SAT) is a standardized test widely used for college admissions, developed and published by the College Board. Our analysis utilizes SAT scores for exam school applicants from two different sources. For the school years 2005 to 2006, SAT data is provided by BPS and includes the students' BPS identifiers and raw scores for the Math, Reading, and Writing sections. We use a crosswalk to link the BPS identifiers to SASIDs. For the years 2007 to 2020, SAT data is provided by DESE and includes the students' SASIDs and raw scores. Both BPS and DESE obtain SAT data for exam applicants from the College Board. These scores are matched to our exam school applicants' file by SASID, using the crosswalk for the earlier years.

If a student takes the SAT more than once, we use the maximum total score (Math + Reading) they achieved among all the exams. SAT score scales were redesigned in 2016. We therefore scale pre-redesign scores using the concordance provided by the College Board.[30] Raw scores are standardized to have mean zero and standard deviation one within a subject-year among exam school applicants in our cohorts who are tested in a given year.

PSAT Data

The Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT) is a standardized test administered by the College Board and co-sponsored by the National Merit Scholarship Corporation (NMSC). All PSAT data were provided by Boston Public Schools (BPS). For the earlier years (2004 to 2005), students are identified by their BPS identifiers, and we use a crosswalk to connect these identifiers to SASIDs. For the years 2006 to 2019, BPS files include SASIDs, so no conversion is needed. PSAT data for the later years also includes scores only for exam school applicants, while the data for the earlier years includes scores for all BPS students, including those applying to exam schools. Since our analysis focuses on exam school applicants, we matched the data appropriately for all the years.

If a student takes the PSAT multiple times, the grade 11 score is given priority if it is available. PSAT score scales were redesigned in 2016. We therefore scale pre-redesign scores using the the conversion scale provided by BPS and the College Board.[31] Raw scores are standardized to have

[30] See Concordance Tables from the College Board

[31] Available at: https://studylib.net/doc/18228722/psat-nmsqt-understanding-scores-2015

mean zero and standard deviation one within a subject-year among exam school applicants in our cohorts who are tested in a given year.

AP Data

Advanced Placement (AP) examinations, offered by the College Board, are taken each May by students as the culmination of year-long AP courses typically offered at the high school level. The AP data in our analysis comes from two different sources. For 2006 and 2007, AP data is provided by Boston Public Schools (BPS) and includes the students' BPS identifiers along with their scores. We use a crosswalk file to link the BPS identifiers to SASIDs for those years. For 2007 to 2020, AP data is provided by the DESE and includes students' SASIDs and AP scores.

Exam school applicants take AP exams across a range of subjects. The tests with 500+ or more takers are Calculus AB, Statistics, Biology, Chemistry, Physics B, English Language and Composition, English Literature and Composition, European History, US Government and Politics, US History, Microeconomics, Macroeconomics, and Spanish Language. The other tests are Art History, Art: Drawing, Art: 2D Design, Art: 3D design, Chinese Language and Culture, Computer Science A, Computer Science AB, Environmental Science, French Language, French Literature, German Language, Comparative Government and Politics, Latin: Vergil, Latin: Literature, Calculus BC, Music Theory, Physics C: Mechanics, Physics C: Electricity and Magnetism, Spanish Literature, and World History.

NSC Data

Data on college enrollment is sourced from the National Student Clearinghouse (NSC) database, which captures enrollment for 96% of U.S. college students (Hindley and Eaton 2018). The Massachusetts Department of Elementary and Secondary Education (DESE) regularly submits the names and dates of birth of in-state high school students for the NSC to search, typically conducting one search each year for high school graduates and one every two years for non-graduates. The resulting dataset includes each student-school enrollment record found in the NSC database, with these records being linked to SASIDs. We use NSC records from the 2007-08 through 2019-20 school years to capture college enrollment for 7th grade exam school applicants in our sample.

Measuring Distance

In our analysis, we utilize measures of school proximity based on distance. For exam school applicants, proximity is determined by the residential coordinates, which are approximated using the centroid coordinates of residential geocodes in Boston. Additionally, the coordinates of Boston exam schools are derived from their addresses. These geocodes are then used to calculate the shortest distance between the applicants' addresses and the exam schools, specifically Boston Latin School and Boston Latin Academy, using the Stata package. `geodist`. This package calculates geodetic distances, representing the shortest paths between two points on the surface of a mathematical model of the Earth.