Blueprint Labs

Discussion Paper #2025.05

Optimal School System and Curriculum Design: Theory and Evidence

Glenn Ellison
Parag Pathak

August 2025

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

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

Optimal School System and Curriculum Design: Theory and Evidence*

Glenn Ellison and Parag A. Pathak†

June 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 student characteristics. Our framework provides a novel interpretation of studies that find zero achievement effects at selective school admission cutoffs. We show that such findings may be consistent with highly effective schools implementing optimal curricula, rather than necessarily indicating ineffective schools. The interpretation depends on the alignment between measured outcome skills and skills emphasized in the curriculum. We test several model predictions using data from a prominent exam school and find supporting evidence that would be difficult to rationalize if selective schools were 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 and Niamh McLoughlin provided invaluable administrative support. We are grateful to partners at a large urban school district who made this study possible. We also benefitted from comments from participants at the Spring 2025 NBER Education Meetings.

†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 skills (see, e.g., Cunha and Heckman (2007), Jackson (2018), and Deming (2023)). This work has important implications for the evaluation of K-12 school effectiveness, particularly when relying solely on standardized test scores. Studies of elite exam schools illustrate this challenge. Abdulkadiroğlu, Angrist, and Pathak (2014) and Dobbie and Fryer (2014) use regression discontinuity designs to find that exam schools in Boston and New York do not increase standardized test scores for students who clear admission cutoffs.[1] However, these schools offer distinct educational experiences. For instance, Boston Latin School requires four years of Latin study, and exam schools generally provide more advanced coursework options than traditional schools. Since state assessments neither test Latin proficiency nor typically evaluate advanced material, they may miss important aspects of what these schools teach. This disconnect between taught and tested skills raises fundamental questions about how to interpret studies of selective schools and, more broadly, how the multidimensional nature of education affects the measurement of school effectiveness.

This paper develops a theoretical framework for understanding how schools optimize their curriculum design, with a particular focus on selective schools. We model curriculum design as a time-allocation problem: schools must decide how to distribute teaching time across different skills based on their student population. Selective schools deploy curricula optimized for the background and ability of their students. The framework offers a new interpretation of research on the effectiveness of selective schools. While a traditional interpretation of the finding of no achievement differences between barely-admitted and barely-rejected students is ineffective schools, our model demonstrates that such findings can actually be consistent with schools making optimal choices about both student selection and curriculum design. Measured school effects depend crucially on which skills are being tested versus which skills the school emphasizes in its curriculum. When schools optimize their curriculum for their student population, they may prioritize skills that aren't captured by standard achievement measures. To test this theory, we examine data from a prominent exam school. We find evidence supporting several novel predictions of our model, patterns that would be difficult to explain if selective schools had no effect on student learning.

Section 2 introduces the paper's theoretical framework with a model where a school system makes

[1] Abdulkadiroğlu, Angrist, and Pathak (2014) also examine PSAT, SAT, and AP scores and both Abdulkadiroğlu, Angrist, and Pathak (2014) and Dobbie and Fryer (2014) look at college-going. We revisit the effects on these outcomes with additional data.

two key choices: how to design curricula at two different schools, and how to allocate students between them. In this model, curriculum design means deciding which type of student each school will serve most effectively. This simple framework leads to an important finding about the limitations of identifying how schools produce educational outcomes. Specifically, when we look at achievement effects for students right at the admissions cutoff, we cannot distinguish between two different scenarios: one where the selective school has an optimally designed curriculum that significantly benefits students, and another where the selective school has no meaningful impact.

Since we can't distinguish between effective and ineffective schools just by looking at achievement at the admission cutoff, we explore a richer analysis and derive several new testable implications of the curriculum-matching model. First, while achievement levels might be continuous at the admission cutoff, the relationship between admission test scores and outcomes should show a change in *slope* at this cutoff. Second, the model predicts that all students, even those performing well at less selective schools, would maximize their expected achievement by stating a preference for the more selective school. We also examine how peer effects influence these predictions. We show that any findings about peer effects from these studies involve tests about both their existence and their particular functional form.

Section 3 develops a more detailed model that views curriculum design as a problem of time allocation. Schools must choose how to divide their limited instructional time across a continuous range of skills, and their optimal allocation depends on their student population. This model serves two purposes: it provides theoretical foundations for the simpler curriculum framework presented in Section 2, and it generates more specific predictions about how school effects should appear in test scores. The observation that an optimally designed system has no discontinuity in achievement at the cutoff requires that the test be perfectly aligned with the value that the designer places on mastery of each skill. With imperfect alignment, there will typically be an upward or downward jump. This leads to two additional testable predictions: first, there should be a discontinuity in the gap between students' performance on basic versus advanced material at the cutoff; second, there should be discontinuities at the admission cutoff when examining performance on individual questions.

Section 4 tests our model's predictions using data from Boston Latin School (BLS), Boston's most prestigious exam school. We analyze seventeen cohorts of students who applied to BLS for 7th-grade admission. Our outcome measures include detailed data on student performance: scores for each individual question on the Grade 10 MCAS (Massachusetts' mandatory high school graduation exam during our sample period), scores on both PSAT and SAT exams, AP exam results, and comprehen-

sive college trajectory data including enrollment, persistence, and graduation. This analysis builds upon Abdulkadiroğlu, Angrist, and Pathak (2014) in two key ways: by adding seven more years of application data to better measure long-term outcomes, and by specifically examining only applicants for 7th grade to the most selective exam school. This focused approach provides a clearer view of how curriculum matching affects students, which may be most relevant at highly selective schools. Our first result confirms the Abdulkadiroğlu, Angrist, and Pathak (2014) finding that students just above the BLS admission cutoff show no overall improvement on Grade 10 MCAS scores compared to students just below the cutoff in our expanded dataset.

The empirical analysis examines five distinct predictions from the curriculum-matching model developed in Sections 2 and 3. First, BLS admission increases performance on SAT English, an important college preparatory assessment. The data also shows changes in the slope of how admission test scores relate to PSAT and SAT performance for English. Second, student application patterns align with the model's predictions about rational choice: the probability of preferring BLS over the next-best school increases with entrance exam scores, with more than half of students preferring BLS even when scoring below the cutoff. Third, examining individual MCAS questions reveals discontinuities in student performance at the admission cutoff, exactly as the model predicts. Fourth, when comparing advanced versus basic test performance, the gap between SAT/PSAT scores and MCAS scores exhibits significant jumps at the cutoff for both Math and English. Finally, looking at measures that capture BLS's advanced curriculum directly, admission increases measures of both AP test participation and performance. These findings collectively provide support for the curriculum-matching model and demonstrate that while BLS has no effect on MCAS scores, it does impact more advanced academic outcomes.

This paper connects to two main areas of research. The first concerns educational settings where curriculum and match effects plays a crucial role. This includes studies of selective schools, tracking, and gifted programs by Duflo, Dupas, and Kremer (2011), Bui, Craig, and Imberman (2014), Pop-Eleches and Urquiola (2013), Card and Giuliano (2016), Cohodes (2020), Bau (2022), Card and Giuliano (2025) and Aitken, Gray-Lobe, Joshi, Kremer, de Laat, and Wong (2025). Related work examines the effects of selective universities (see, e.g., Dale and Krueger (2002) and Mountjoy and Hickman (2021)). Three studies using regression discontinuity designs find positive earnings effects from attending selective public universities (Hoekstra (2009), Zimmerman (2014), and Bleemer (2024)). Chetty, Deming, and Friedman (2023) examine selective private universities, focusing on non-traditional outcomes like elite graduate school attendance and public service leadership positions,

outcomes specifically targeted by these institutions.[2] Their emphasis on measuring outcomes that match institutional goals aligns closely with our approach. The importance of curriculum also appears in research on school accountability. For instance, Jacob (2005) shows that Chicago's school accountability system improved performance on skills tested by high-stakes exams but not on other assessments, while Cohodes (2016) looks at responses on specific test questions to study whether Boston charter schools are more effective on commonly tested standards compared to infrequently tested standards. Our study also uses item-level test responses to measure effects on specific skills.

The paper also contributes to the economics of education literature that uses microeconomic theory models to understand empirical findings. For example, Lazear (2001) explains class size effects through a model where student disruption affects learning, while Urquiola and Verhoogen (2009) shows how school and household choices about class size can invalidate regression discontinuity designs by creating discontinuities in family income at cutoffs. Our work follows this tradition by showing how curriculum design can explain why selective schools might show no achievement effects at admission cutoffs despite being effective. We then develop and test specific predictions to distinguish between this explanation and the simpler interpretation that these schools are ineffective.

2 A Curriculum-Matching Model of School System Design

This section develops a model of how school systems can be optimally designed. The model has two key elements: first, an agreed-upon outcome ("achievement") that depends on both individual learning style and school curriculum; second, schools observe an indicator of each student's learning style. The school system then makes two choices to maximize total expected outcomes: it decides which students attend which schools, and it chooses each school's curriculum.

2.1 Model of School System with Selective Schools

Consider a school system with two schools serving a continuous distribution of students. Each student has a learning style θ ("type") that, after normalization, can be represented as their percentile in the uniform distribution on $[0, 1]$. The school system observes a signal r_i for each student i's learning style. This signal r has a continuous distribution with full support on interval R. The signal satisfies the monotone likelihood ratio property (meaning higher signals suggest higher learning styles), and

[2] 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."

has conditional density $g(\theta|r)$ that varies smoothly with r.[3]

A school system makes two key design choices: 1) an assignment, denoted by function $A : R \to \{1, 2\}$ that maps each student's signal r_i to either school 1 or 2, and 2) a curriculum choice, c_1 and c_2, for each school. Let the achievement of a student i of type θ_i in a school s with curriculum c_s be given by

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

where $h(\theta_i)$ represents the student's underlying ability (assumed to be differentiable in θ), $m(\theta_i, c_s)$ captures how well the curriculum matches the student's learning style, τ_s represents school-specific effects that apply equally to all students, and ϵ_i represents random individual shocks independent of other factors. The curriculum match function $m(\theta, c)$ is smooth and exhibits strictly increasing differences with no single curriculum working best for all students. A simple example is $m(\theta, c) = -(\theta - c)^2$, where achievement decreases with the square of the distance between student type and curriculum. School effects τ_s can differ between schools, but not so dramatically that putting all students in one school becomes optimal.

The school system aims to maximize average student achievement by choosing both curricula and student assignments. The model makes two important simplifying assumptions: each school must offer just one curriculum, with no costs for curriculum choices, and schools can accommodate any distribution of students without capacity constraints or crowding effects.

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 $c_2 = \arg\max_c m(\theta', c)$.

[3]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.

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 first-order stochastic dominance 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 Admissions 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 $y_s(r) \equiv \mathbb{E}_{g(\theta|r)}[h(\theta) + m(\theta, c_s)] + \tau_s$. 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 logic behind this result is both intuitive and broadly applicable: if achievement shows a jump at the cutoff, the school system could improve overall outcomes by simply reassigning students just above or below the cutoff to the other school. However, this logic depends on the assumption that schools can accept any number of students. The result might not hold if, for example, school 1 faced capacity constraints that limited its enrollment.

Proposition 2 shows that when curricula and student assignments are optimally designed, regression discontinuity estimates will find zero effect of attending a selective school. However, this same zero effect could also emerge from what we call the **zero value-added model** of selective schools, where schools are identical ($\tau_1 = \tau_2$) and curriculum matching doesn't matter ($m(\theta, c) = 0$ for all θ and c). We collect testable implications 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.*

Finding no achievement effect at the admission cutoff creates a fundamental identification challenge because this result is consistent with two very different models of education. In the zero value-added model, school assignment doesn't matter. It affects neither individual nor aggregate student achievement. In contrast, under the curriculum-matching model, school assignment impacts both individual and total achievement. While these models produce identical predictions about achievement at the cutoff, they can be distinguished through other means. The following analysis presents several testable predictions that emerge from examining the broader relationship between school assignment and various measures of achievement.

2.4 Slope Changes at Admissions Cutoffs

While the curriculum-matching model predicts no jump in achievement levels at the admission cutoff, it does predict other discontinuities. Specifically, because schools offer different curricula across the cutoff, the model predicts a sudden change in how student achievement responds to ability - that is, a discontinuity 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. ■

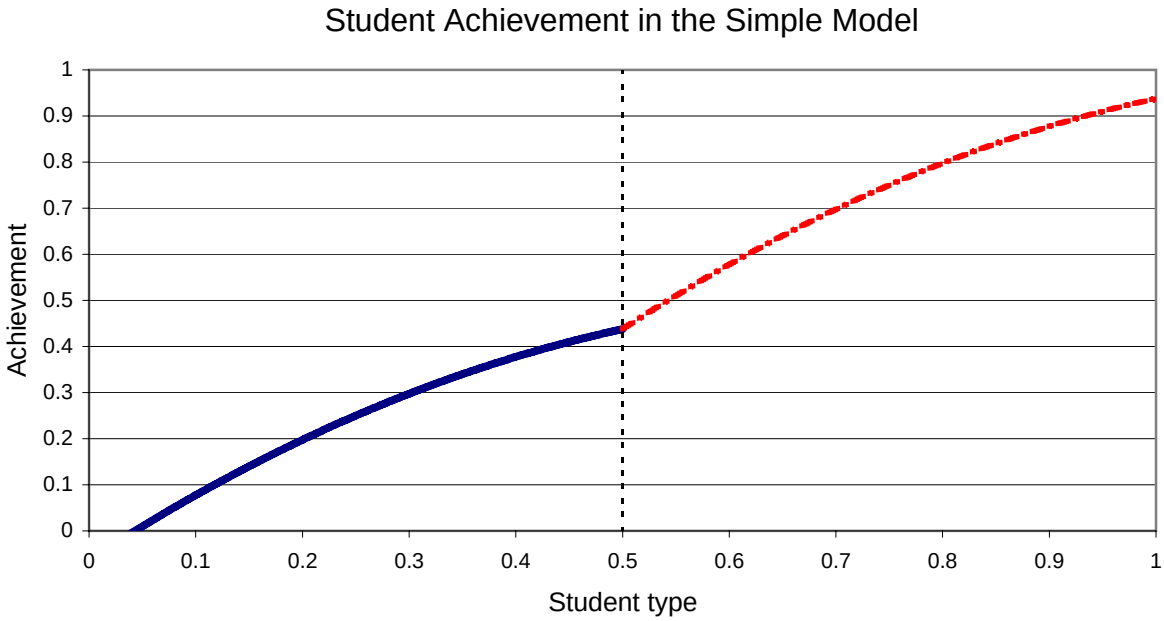
To understand why the slope changes at the admission cutoff, consider the simple case where the signal r perfectly indicates learning style θ. Near the cutoff $\hat{r}$, student achievement varies with r for two reasons. First, it varies through the school-independent component $h(r)$, which changes smoothly across the cutoff. Second, it varies through the curriculum match $m(r, c_s(r))$, which behaves differently on each side. Below $\hat{r}$, students face a curriculum in school 2 that's too basic for them, so the match quality decreases as r rises. Above $\hat{r}$, students face a more advanced curriculum in school 1, so the match quality improves as r rises. This creates a discontinuity in the slope at the cutoff.

While Proposition 3 formally describes the difference in left and right derivatives at the cutoff point $\hat{r}$, the logic suggests this slope difference extends beyond just that point. Over a broader range below $\hat{r}$, students face increasing mismatch with the basic curriculum, while above $\hat{r}$, students become better matched to the advanced curriculum. Therefore, unless the relationship between ability and achievement ($\mathbb{E}_{g(\theta|r)}h(\theta)$) is highly nonlinear near $\hat{r}$, we should observe a consistently lower slope below the cutoff than above it.

A simple numerical example illustrates how slopes change in the model. Consider these assumptions: student types θ are uniformly distributed on [0,1], the signal perfectly reveals type ($r = \theta$), and curriculum mismatch creates quadratic losses: $m(\theta, c) = -(\theta - c)^2$. Under these conditions, the optimal design features a cutoff at $\hat{\theta} = \frac{1}{2}$ and school curricula at $c_1 = \frac{1}{4}$ and $c_2 = \frac{3}{4}$. In the graph, the bold blue line shows achievement for students in school 2 (left side), while the red dashed line shows achievement for school 1 students (right side). While achievement levels remain continuous at the cutoff, the slope increases sharply there. Moreover, because mismatch costs are convex (making match quality concave), any non-local estimate of the slope change would actually underestimate the true jump in derivative at the cutoff due to match quality effects.

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 slope discontinuity provides a key way to distinguish between the curriculum-matching and zero value-added models. In the zero value-added model, there should be no sudden change in slope



at the cutoff, as long as $\mathbb{E}_{g(\theta|r)}h(\theta)$ is differentiable at $\hat{r}$.[4]

2.5 Application Behavior at Admissions Cutoffs

Abdulkadiroğlu, Angrist, and Pathak (2014) highlight an apparent paradox: parents strongly desire to send their children to selective schools, yet there's no jump in student outcomes at the admission cutoff. They suggest two possible explanations: either parents wrongly assume that high-achieving peers indicate high school quality, or they value these schools for reasons unrelated to academic achievement.

The curriculum-matching model suggests another explanation for strong parent demand: while there might be no benefit right at the admission cutoff, students above the cutoff could still gain substantially from attending the selective school. Furthermore, when parents apply, they face uncertainty about their child's learning style (θ) and test performance (r). In this situation, stating a preference

[4]This raises a practical question about estimating slope changes at cutoffs. If $\mathbb{E}_{g(\theta|r)}h(\theta)$ is convex around $\hat{r}$, a highly non-linear relationship might be mistaken for an upward jump in slope. Such convexity could occur if each percentile increase in r yields progressively larger increases in future test scores - as would happen if r perfectly predicted normally distributed test scores. Conversely, if the entrance exam becomes less discriminating among high achievers, it would create a concave relationship, making it harder to detect a positive slope change. However, in our application, where the cutoff isn't far in the tail of the distribution, such highly nonlinear relationships seem unlikely in small windows around the cutoff.

for the selective school is rational because if the choice becomes relevant in the sense that their child scores well enough for admission, then the student will be above the cutoff and thus benefit from the school's curriculum.

Consider extending the curriculum-matching model to include student choice. Students observe a noisy signal $t_i = r_i + \xi_i$ of how the school system will assess them, where ξ_i is a random variable distributed independently of (θ_i, r_i). Each student's utility from attending school s is given by $u_i = y(\theta_i, c_s) + \nu_{is}$, where the ν_{is} are iid preference shocks. Students must submit their school preferences after seeing t_i but before knowing their true θ_i or r_i. Under these conditions, the model predicts that every student, even those likely to achieve higher scores in school 2, will prefer school 1 with probability of 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 ranking 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 iid preference shocks, $\nu_{1s} - \nu_{2s}$ is symmetrically distributed around zero, so the probability that school 1 is the utility-maximizing choice is greater than one-half. ■

In this model, students with higher r values are typically more likely to prefer school 1. This occurs because higher r values tend to produce higher observed signals t, and students who see higher t values estimate a higher r and thus greater expected benefit from school 1. However, the relationship between r and school preference isn't as straightforward as the previous results. While a first-order stochastic dominant increase in r raises the probability of admission to school 1, if this increase mainly affects cases where r is just above $\hat{r}$, the expected achievement advantage from attending school 1 (conditional on admission) might not increase monotonically with t. This means we can construct distributions where the probability of preferring school 1 doesn't monotonically increase with r.

Consider a simple example that illustrates the key insights while avoiding pathological cases. Assume that θ is uniform on [0,1], schools have equal base effects ($\tau_1 = \tau_2$), base achievement equals type ($h(\theta) = \theta$), curriculum mismatch has quadratic costs ($m(\theta, c) = -(m - c)^2$), schools perfectly observe type ($r_i = \theta_i$), signal noise ξ is uniform on $[-\sigma, \sigma]$, relative school preferences $\nu_{i1} - \nu_{i2}$ are uniform on $[-\delta, \delta]$, signal noise is moderate ($\sigma < 0.25$), and preference shocks are at least as large as signal noise ($\delta \geq \sigma$).

In this model, students who end up exactly at the cutoff choose school 1 with probability strictly above one-half: $p(\hat{r}) = \frac{1}{2} + \frac{\sigma}{4\delta}$. To understand why, consider what happens when the cutoff is $\frac{1}{2}$ (as in the uniform model). Students who later score at the cutoff initially saw different signals. Those who saw the lowest possible signal ($\theta' = \hat{\theta} - \sigma$) knew their choice would only matter if their true type was exactly $\hat{\theta}$. At this point, achievement would be equal at both schools, so exactly half choose each school. Students who saw signals between $\hat{\theta} - \sigma$ and $\hat{\theta} + \sigma$ knew their true type would be uniformly distributed on $[\hat{\theta}, \theta' + \sigma]$.[5] This knowledge leads them to prefer school 1 with probability greater than one-half. For instance, students who saw the highest possible signal ($\hat{\theta} + \sigma$) choose school 1 with probability $\frac{1}{2} + \frac{\sigma}{2\delta}$ because their expected achievement gain is σ. The probability varies linearly between these extremes, yielding the formula above.

In this model, we can prove several properties about $p(r)$, the probability of choosing school 1: it weakly increases with r, its derivative at $\hat{r}$ is $\frac{1}{4\delta}$, and it equals 1 for $r > \hat{r} + \sigma + \delta$. The graph shows these probabilities calculated with $\sigma = 0.15$ and $\delta = 0.15$. Under these parameters, the probability of preferring school 1 rises smoothly from one-half for very low-type students to one for very high-type students, reaching about three-quarters for students near the cutoff.[6]

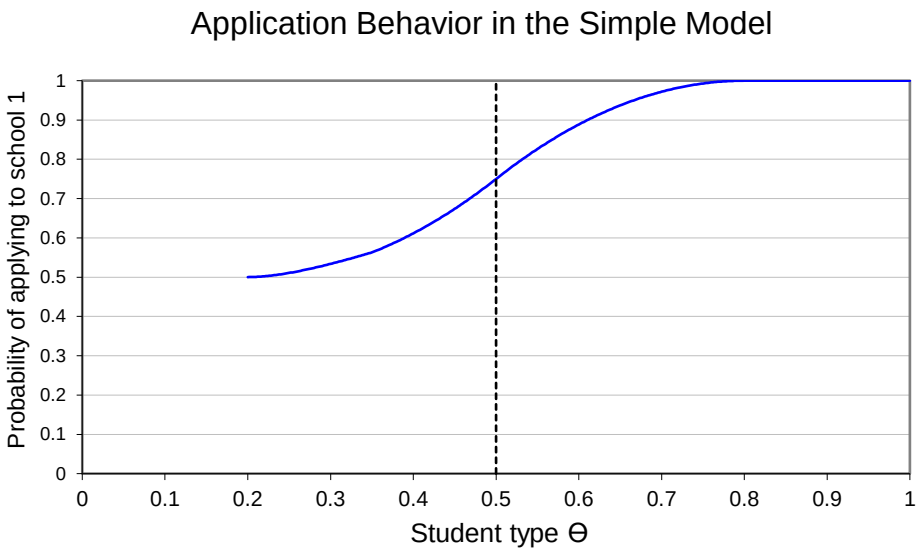
PREDICTION 2: *In the curriculum-matching model, if parents receive an imperfect signal about their children's types and have independent identically distributed idiosyncratic preferences about schools, then the probability of ranking the top school first will be strictly greater than one-half for students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

2.6 Peer Effects in Education Production

The model developed so far assumes student achievement depends on three factors: individual learning style (θ_i), school characteristics, and curriculum. This section explores how the addition of peer effects alters our results.

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

[6] The model reaches exact probabilities of $\frac{1}{2}$ and 1 (rather than approaching these values asymptotically) because ξ and η have finite supports. For instance, students with sufficiently high types will receive signals indicating that their gain from the best school certainly exceeds any possible taste shock. The finite support also leaves choice probabilities undefined for students whose signals make admission to the selective school impossible. For the graph, we assume these students choose school 1 with probability one-half, effectively treating them as if they believe their type would be $\hat{r}$ (the minimum for admission) if admitted.



Suppose we modify our achievement model to include peer effects:

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

where $b > 0$ measures peer effect strength, and $w_{ji} \geq 0$ represents student j's influence on student i. We make two key assumptions in this formulation. First, each student's total influence on others remains constant ($\sum_k w_{jk} = 1$ for all j), so school assignment can only redirect this influence, not change its magnitude. Second, students only influence peers at their own school ($w_{jk} = 0$ if students j and k attend different schools). To simplify the discussion below, we assume that schools observe θ perfectly. 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.*

Pop-Eleches and Urquiola (2013), Abdulkadiroğlu, Angrist, and Pathak (2014) and Dobbie and Fryer (2014) suggest interpreting regression discontinuity estimates of exam school effects as measuring the impact of exposure to different peer groups. However, Proposition 5 shows that the size of

this effect depends crucially on how we model peer influence. Under the common assumption that all students in a school affect each other equally, the discontinuity in peer effects at the cutoff is $b(\mathbb{E}[\theta|\theta > \hat{\theta}] - \mathbb{E}[\theta|\theta < \hat{\theta}])$, where b represents the strength of peer effects.

At the other extreme, consider a scenario where students only interact with peers who are most similar to themselves. Specifically, each student is influenced by exactly one other student: the one whose characteristics are most closely matched to their own. In mathematical terms, for any two students j and k, $w_{jk} = 1$ if student θ_j is closer to θ_k than is $\theta_{k'}$ for any $k' \notin \{k, j\}$, and $w_{jk} = 0$ otherwise.[7] In this case, there would be no discontinuity at the cutoff point. The change in the peer effect at the cutoff would be equal to b multiplied by the difference between the second-lowest value above the cutoff and the second-highest value below the cutoff. This difference would be similar in magnitude to the difference between the students closest to either side of the cutoff and an RD analysis would not detect any jump at the cutoff.

As a result, any conclusions about peer effects drawn from RD studies of selective schools must be interpreted carefully. These findings actually test two things simultaneously. They test whether peer effects exist, but only under specific assumptions about how peers influence each other. This is particularly evident in the curriculum-matching model. Without first specifying exactly the functional form of how peers affect one another, we cannot interpret what the sudden changes at cutoff points tell us about peer effects in general.

3 A Model with Explicit Curriculum Design

Our discussion so far has treated curriculum design in a simplified way: schools simply pick a single number that represents their curriculum. In our earlier example with a quadratic mismatch function, this number could be thought of as identifying the type of student the curriculum serves best, with other students' outcomes depending on how far their abilities differ from this ideal student. We will now develop a more detailed and realistic model of curriculum design. In this new model, schools teach multiple skills and must decide both which skills to teach (the extensive margin) and how much time to devote to each skill (the intensive margin). This expanded framework allows us to examine how different types of assessments might measure student performance differently. Additionally, this more detailed model provides theoretical support for the simpler approach we used in the previous

[7] 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.

section.

3.1 A Model of Skill Accumulation

Consider a model in which achievement is measured by how many skills a student acquires from a range of possible skills, labeled from $[0, 1]$. Learning happens probabilistically during instruction: when a school spends time dt teaching a skill x at unit intensity, a student with ability level θ has a probability of $\theta x dt$ of mastering that skill (assuming they don't already know it). Additional teaching of an already-mastered skill offers no benefit. Skills with higher x values are easier to learn, and students with higher θ can be thought of as higher ability students who will (in expectation) learn any skill more quickly.

When a skill x is taught for a total duration t, the probability that a student of ability θ will learn that skill is given by the formula

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

While our model treats learning as an all-or-nothing event (either knowing or not knowing the skill), this formula can also be interpreted as describing how students gradually master skills over time, with a student's mastery increasing smoothly in the time devoted to the skill. The exponential structure naturally captures an important learning principle: the marginal benefit of additional instruction time decreases as more time is spent teaching a skill.

3.2 Curriculum Design as a Time-Allocation Problem

The model treats curriculum design as a decision about time allocation: for each skill x, we must choose how much teaching time $t(x)$ to devote to it. When a student of ability level θ is taught under a particular curriculum, their achievement is calculated by taking a weighted average across all skills. Each skill's contribution to this average is weighted by its importance $v(x)$. The formula for expected achievement is:

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

This formulation builds on our previous model by adding multiple dimensions to curriculum design, while preserving the core property that each student ability level θ has an optimal curriculum, and student achievement falls when the curriculum moves away from this optimum.

We first consider a single school teaching a group of students whose ability levels θ are distributed

according to a density function $g(\theta)$ over a set Θ. The school has a fixed amount of total instructional time T and must decide how to allocate this time across different skills to maximize overall student achievement. This optimization problem can be written as:

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

Here, the school tries to maximize the average achievement across all its students (the first equation) while ensuring that the total time spent teaching all skills equals the available time T (the constraint).

For any curriculum design $t(x)$, we can calculate the benefit of adding a little more teaching time to skill x. This marginal benefit is given by:

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

When we start teaching a new skill ($t(x) = 0$), the initial marginal benefit is:

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

This marginal benefit gets smaller as we spend more time teaching the skill, and eventually approaches zero as teaching time becomes very large ($t(x) \to \infty$). This pattern of diminishing returns helps us characterize 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

We now solve for the optimal teaching time $t^*(x)$ in a simplified case in which the school values all skills equally ($v(x) = 1$) and teaches students of the same ability level. By examining numerical solutions, we can illustrate how the optimal curriculum changes based on the students' ability level and the total available instruction time.

When all students have the same ability level θ, the condition that marginal benefits are equalized across skills for any skill that is taught ($t^*(x) > 0$) implies:

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

Solving for the optimal teaching time gives us:

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

This formula shows how the relationship between teaching time and skill difficulty depends on λ. When $\lambda > \theta/e$, more difficult skills receive more teaching time. However, when λ is smaller, teaching time is non-monotone. It peaks at skill level $\frac{e\lambda}{\theta}$ and then declines. Since λ equals θ when total teaching time T is very small and approaches 0 as T becomes very large, this means that schools with limited time will focus more on harder skills, while schools with more time will have a more balanced distribution of teaching time across skill levels.

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:

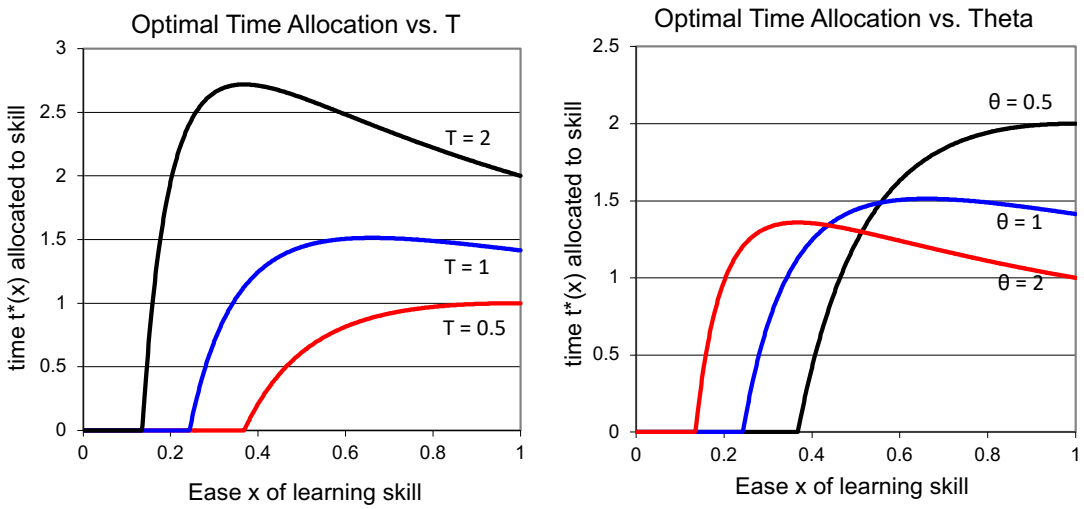
$$t^*(x) = \frac{\log(x) + \sqrt{2\theta T}}{\theta x} \quad \text{for } x \in \left[e^{-\sqrt{2\theta T}}, 1\right].$$

The graph on the left below demonstrates how optimal teaching time $t^*(x)$ changes with total available time T. When teaching time is limited ($T = 0.5$), shown by the lowest curve, schools spend progressively more time on harder skills. However, as more total time becomes available ($T = 1$ and $T = 2$), shown by the higher curves, two changes occur. Schools teach a wider range of skills, and they no longer always devote more time to harder skills. Instead, they develop a more balanced curriculum where moderate-difficulty skills might receive the most instruction time.

The graph on the right below illustrates how optimal teaching time $t^*(x)$ varies with student ability θ when total teaching time is fixed at $T = 1$. For lower-ability students ($\theta = 0.5$), shown by the rightmost curve, the optimal curriculum focuses progressively more time on easier skills. However, as student ability increases ($\theta = 1$ and $\theta = 2$), the optimal curriculum changes in two key ways: it includes a broader range of skills, extending into more difficult skills, and it redistributes time away from the easiest skills toward harder ones. The differences are most dramatic at the extremes. For the easiest skills (those with x close to 1), the curriculum for high-ability students ($\theta = 2$) spends only half as much time as the curriculum for lower-ability students ($\theta = 0.5$). Notably, the skills that receive the most attention in the high-ability curriculum are skills that the lower-ability curriculum considers too difficult to teach at all.

3.4 Explicit Curriculum Design and Student Achievement

Consider now a system with two schools where school 1 enrolls higher-ability students ($\theta \in [\hat{\theta}, 1]$) and school 2 enrolls lower-ability students ($\theta \in [0, \hat{\theta}]$). Let $g_i(\theta)$ represent how student abilities are distributed within each school i. School time allocations will have a single-crossing property. School 2 (the lower-ability school) will devote more time to skills that offer high immediate returns (large $xv(x)$), while school 1 (the higher-ability school) will spend more time on skills with lower immediate returns (small $xv(x)$).



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 $t_1^*(x) = t_2^*(x) = 0$.*

Proposition 7 extends our understanding of how schools allocate teaching time across skills of varying difficulty. When all skills are equally valuable, the school with lower-ability students focuses more on basic skills. This happens because higher-ability students master basic skills quickly, leading their school to shift time toward more challenging skills. The proposition generalizes this insight by showing that what matters is not just skill difficulty, but rather the product of difficulty and value ($xv(x)$), which represents the initial benefit of teaching a skill. The complete proof can be found in Appendix A.

An important implication of Proposition 7 follows: except in rare cases where many skills have exactly the same initial benefit ($xv(x) = w$), the two schools will spend different amounts of time on almost every skill that school 1 teaches ($t_1(x) \neq t_2(x)$ for almost all x where $t_1(x) > 0$). This means that which school a student attends directly affects what they learn, even for students right at the cutoff ability level $\hat{\theta}$. Students who attend school 1 (the higher-ability school) will learn more of the skills with low initial benefits ($xv(x) < w$), while those who attend school 2 (the lower-ability school) will learn more of the skills with high initial benefits ($xv(x) > w$). This result 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.*

When researchers use RD designs to study selective schools' effects, they typically measure student achievement through comprehensive exams that test many different skills. Even though students will learn any given skill better at one school than the other, our earlier impossibility result (Proposition 1) still holds in this more detailed curriculum model, provided that the test used to measure achievement aligns with the school system's educational goals in a particular 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 logic behind this proposition is simple. When the test score calculations match the objectives that schools are trying to maximize, we can apply the same reasoning as in our simpler model: an optimal school system cannot show a sudden jump in performance at the admissions cutoff. 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.*

Students typically take many different tests throughout their education. Our model makes specific predictions about their relative performance on these tests. Generally, students at the higher-ability school (school 1) will perform relatively worse on tests of basic skills but better on tests of more advanced material. We can express this pattern more precisely as our next result.

Proposition 9 *Let $q_j(x)$ be the fraction of questions on test j testing skill x. Suppose the school assignment is optimal and test A has more questions on hard skills and fewer on easy skills in the*

sense that $q_A(x) < q_B(x)$ *if and only* $xv(x) < w$ *where* w *is the single crossing point defined in Proposition 7. Then* $\mathbb{E}(y_A(\theta) - y_B(\theta)|\theta)$ *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 helps distinguish our curriculum-matching model from a model where schools add no value. In a zero value-added model, where test scores are given by $y_{ij} = h_j(\theta_i) + \tau_s + \epsilon_{ij}$, the difference between scores on two tests A and B would be $\mathbb{E}[y_{iA} - y_{iB}|\theta] = h_A(\theta) - h_B(\theta)$. This difference would show no sudden jump at the cutoff if performance on each test varies smoothly with student ability. We can test this prediction by comparing two types of scores: SAT/PSAT scores (which focus on more advanced material) and state standardized test scores (which were required for graduation until recently and emphasize more basic skills).

4 Evidence from the Boston Latin School

We now test the predictions of both the curriculum-matching model and the zero value-added model by analyzing data from students who applied to the Boston Latin School (BLS).

Boston Public Schools (BPS) is an appealing setting to test our models for two reasons. First, its exam schools use strict admissions cutoffs, which remained consistent during our study period.[8] Second, BPS provided seventeen years of detailed student data, including application preferences, entrance exam scores, and various outcomes. We focus specifically on Boston Latin School (BLS), the most selective of Boston's exam schools, because curriculum-matching concerns seem most salient there.[9] BLS serves as a pathway to elite colleges and employs a rigorous college preparatory curriculum that differs markedly from curricula designed to maximize MCAS scores.[10] To analyze educational

[8] BPS changed its policy in 2021 to include neighborhood characteristics and GPA (Barry 2021).

[9] BLS significantly outperforms the other two exam schools in terms of test scores. For the 2023-24 school year, BLS students averaged 1304 on the SAT, while students at Boston Latin Academy and O'Bryant High School for Science and Mathematics scored notably lower, averaging 1123 and 1105 respectively.

[10] 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."

outcomes, we examine three assessments: the 10th grade MCAS, 11th grade PSAT, and SAT tests (using the maximum SAT score for students who take it multiple times). These tests are particularly useful because they occur after students have experienced several years of BLS curriculum, test different types of material, and take place during the latter part of high school (grades 10-12). We also examine college enrollment patterns using National Student Clearinghouse data.

Abdulkadiroğlu, Angrist, and Pathak (2014) conducted RD analyses around the admissions cutoffs for all three Boston Exam schools for both 7th and 9th grade applicants. Their pooled estimates averaged potentially different effects across schools and across varying years of exposure to the exam school curriculum. Our dataset extends theirs by seven additional years, providing a larger sample that enables us to investigate more subtle patterns that weren't examined in the earlier paper. These include changes in test score slopes and question-level differences, and the longer-term outcome of college completion. The larger sample also allows us to focus specifically on Boston Latin School, where curriculum matching concerns seem most relevant.[11] Further details about data processing can be found in the data appendix.

Table 1 reports descriptive statistics on the sample. As in many large urban districts, most Boston 6th graders are Black or Hispanic. Boston students have average scores below the state average on MCAS tests and below national averages on the PSAT and SAT. 60% of students for whom we have college attendance data attend college and only 22% graduate from college.[12] Students who apply to BLS, particularly those who list it as their first choice, differ from the general Boston student population: they are less likely to be minorities, achieve higher MCAS scores, and have higher college attendance rates. This pattern becomes even more pronounced when we look at applicants within 20 percentile ranks of the BLS admissions cutoff. In this group, less than half are Black or Hispanic. Their Grade 6 MCAS scores exceed the state average by 1 standard deviation (σ) in Math and 0.7σ in English. (For all MCAS outcomes, we standardize scores to have mean zero and standard deviation one across all Massachusetts test-takers.) These students also score significantly above national averages on the SAT and have much higher rates of college attendance, persistence,

[11] We also limit our analysis to applicants for 7th grade – the entry point for 85-90% of BLS students – ensuring that students in our sample had the same number of years of potential exposure to the BLS curriculum prior to the test we examine.

[12] Our college attendance data only include student who reached 12th grade at a public high school in Massachusetts. There is nontrivial dropout prior to 12th grade in BPS, so these statistics overstate college attendance/graduation in the full BPS population. Dropout is much less common for the students who apply to BLS and score around the cutoff, and Table A1 shows that we do not find evidence for discontinuities in the availability of the data at that cutoff.

and graduation compared to the overall Boston student population.[13] These distinct characteristics emphasize the importance of examining educational outcomes that are particularly relevant for this high-achieving student population.

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

We start by using our expanded dataset to verify a main finding from Abdulkadiroğlu, Angrist, and Pathak (2014). We look for sudden changes in Grade 10 MCAS Math and English scores at the BLS admission cutoff. The MCAS exam was required for high school graduation in Massachusetts during our sample period and is a central metric in the state's accountability system.

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. Figure 1 shows how students' Grade 10 MCAS scores (measured in standard deviations) relate to their ranking on the BLS entrance exam, where ranking is expressed as a percentile among that year's exam school applicants. 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.[14]

The dashed vertical lines Figure 1 mark the BLS admission cutoffs. Like Abdulkadiroğlu, Angrist, and Pathak (2014), we find no sudden changes in either math scores (left panel) or English scores (right panel) at these cutoffs. This finding aligns with both the zero value-added model and certain versions of the curriculum-matching model. In the curriculum-matching framework, this pattern could emerge in two scenarios: either when the school system optimizes its curricula and student assignments to maximize MCAS performance without peer effects, or when the curricula target more advanced material than what appears on the MCAS and peer effects are precisely strong enough to offset what would otherwise be a drop in performance at the cutoff, as we explained in Section 2.6.

4.1 Prediction 1: Changes in Slopes at the Admissions Cutoff

We now test additional predictions of the curriculum-matching model, starting with Prediction 1. We look for changes in the slope of the relationship between achievement and entrance exam scores at the BLS cutoff. This prediction stems from how well students match with their school's curriculum. Below the cutoff, match quality decreases as students become increasingly overqualified for their

[13] We define persistence as attending college for at least four semesters.

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

school's curriculum. Above the cutoff, match quality improves as students become better suited to the more advanced curriculum. This shift from declining to improving match quality should create a discrete upward jump in the derivative of the achievement-entrance score relationship at the cutoff.

Figure 2 shows plots similar to Figure 1, but for four college-preparatory tests: PSAT and SAT math, and PSAT reading and SAT critical reading. We standardize PSAT and SAT scores to have mean zero and standard deviation one within each subject-year, considering only BPS exam applicants. As before, each dot represents the average score for students in one-percentile bins within 20 percentiles of the BLS cutoff, and the curves show local linear regression estimates on either side of the cutoff.

The graphs show evidence of slope changes in the relationship between PSAT and SAT Reading scores and entrance exam percentile rank at the BLS cutoff. However, these figures also highlight the challenges of isolating such slope changes, as they may reflect broader non-linear patterns around admissions cutoffs rather than true changes in the slope. To assess whether these observed slope changes are statistically significant, Table 2 presents regression discontinuity estimates of the slope changes at the cutoffs, along with their 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

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

In this equation, D_i indicates an offer of admissions (that is, if $r_i \geq 0$), $f(r_i)$ and $g(r_i)$ are potentially nonlinear functions of the entrance exam score normalized to have slope one at $r = 0$, and α_t controls for the application year. The coefficient of interest is the change in slope, γ_1, at the admissions cutoff. We use the optimal bandwidth computed from Imbens and Kalyanaraman (2012) (IK) with a tent-shaped edge kernel for these estimates.[15] This specification is a simpler version of the one used in Abdulkadiroğlu, Angrist, and Pathak (2014), as we focus only on applicants for 7th grade who list BLS as their first choice.[16]

Table 2 presents estimates for math and reading scores across three tests: MCAS, PSAT, and SAT.[17] For PSAT and SAT Math outomes, the estimated slope changes are positive, but not statisti-

[15] It's important to note that our analysis differs from studies using regression kink design (see Card, Lee, Pei, and Weber (2015) and Ganong and Jäger (2018)), where the probability of assignment changes slope at a certain point. In our case, there is a sharp discontinuity in assignment at the admissions cutoff.

[16] See page 154 of Abdulkadiroğlu, Angrist, and Pathak (2014). Since we look at only first-choice applicants, we don't need the additional controls for assignment risk that were used in Abdulkadiroğlu, Angrist, Narita, and Pathak (2022).

[17] Appendix Table A1 examines whether students above and below the cutoff differ in how likely they are to have recorded outcomes. We find no differences in data availability for PSAT, SAT, PSAT-MCAS, or SAT-MCAS scores, suggesting that missing data is unlikely to affect our estimates for these outcomes.

cally significant. For the PSAT and SAT English outcomes, the estimated slope changes are positive and statistically significant. which support Prediction 1 of the curriculum-matching model. Specifically, we find significant slope changes at the 5% level for PSAT English and at the 1% level for SAT English. These findings are particularly compelling given how challenging it is to estimate slope changes in an RD framework.[18]

The traditional model of peer effects, where each student receives the same additive benefit based on average student quality, wouldn't create a slope change at the cutoff. Similarly, our alternative model from equation (2), where benefits come from the most similar peer, produces neither a discontinuity nor a slope change at the cutoff. However, other peer effect models could generate slope changes. Consider a model where a student of type θ at school s benefits from higher average peer quality but is penalized for being different from their peers, expressed as $\alpha_0\overline{\theta}_s - \alpha_1(\theta - \overline{\theta}_s)^2$. This formulation closely resembles our curriculum-matching model and could produce a slope increase at the cutoff through the same mechanism.

4.2 Prediction 2: Stated Applicant Preferences

Section 2.5 showed how the curriculum-matching model helps explain why parents strongly prefer selective schools despite limited evidence of benefits for students near the admissions cutoff. According to Prediction 2, school preferences for a selective school are only relevant if the student turns out to be above the cutoff, in which case benefits do exist. If purely idiosyncratic preferences are symmetrically distributed, a majority of students at every entrance exam score level (even those who would actually be better matched to the less selective school) should state a preference for the more selective school. Additionally, the proportion of students preferring the more selective school should increase with entrance exam scores.

To test this prediction, we estimate how entrance exam scores affect the probability that students rank BLS above Boston Latin Academy (BLA), the second-most selective exam school, while accounting for how far students live from each school. Our analysis includes all students who applied to either BLS or BLA, not just those near the BLS cutoff. This broader sample includes many students with

[18]Note that our results are derived from local linear regressions. Standard errors of the slope changes estimated via Calonico, Cattaneo, and Titiunik (2014) are much larger than the effects we estimate. We believe our approach is more appropriate for two reasons. First, under the zero value-added hypothesis, the second derivative of $\mathbb{E}(y|r)$ is identical on both sides of the cutoff, so our theory suggests the bias their method corrects for should be zero. Second, since our model predicts a consistent difference between slopes on either side of the cutoff (not just a change at the cutoff point), we aren't concerned that the IK bandwidths consider a substantial interval around the cutoff.

scores well below the cutoff who weren't included in our previous figures that focused only around the cutoff. Let p_{it} equal 1 if applicant i in cohort t ranked BLS over BLA. We count a student as preferring BLS if they either ranked BLS higher than BLA or included only BLS on their preference list. We estimate this relationship using:

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

Here, μ_t represents cohort fixed effects, d_{it}^s measures the distance between the student's home and school $s \in \{\text{BLS}, \text{BLA}\}$ with coefficient β^s, and $g(\cdot)$ is a flexible function of the admission test score.[19] While we include controls for distance since it might affect school preferences, the patterns remain almost identical without these controls.

Figure 3 plots the local-linear fit of this equation, which align with the curriculum-matching model's predictions for students who have imperfect information about their match quality. More than 50% of applicants prefer BLS over BLA at every entrance exam score level, even among the lowest-scoring students. Around the BLS cutoff, about 80% prefer BLS, with this percentage increasing across a wide range of nearby scores. Among the highest-scoring students, about 90% prefer BLS. These are rational choices in the curriculum-matching model: students near the cutoff expect they'll be better matched at BLS *if admitted* (though this expectation may prove wrong ex post), while top-scoring students (correctly) anticipate they'll be much better matched at BLS.

As we discussed in the theory section, it's hard to explain these strong preferences for BLS using a zero value-added model with fully rational choice. Abdulkadiroğlu, Angrist, and Pathak (2014) suggest parents may be operating under an "elite illusion," mistakenly believing in value-added that doesn't exist. However, this explanation requires either that higher-scoring students (or their parents) make larger errors in their beliefs, or that unobserved characteristics correlate with both entrance exam performance and BLS preference relative to BLA.

4.3 Prediction 3: Discontinuities in Question-Specific Performance

The curriculum-matching model with explicit curriculum design predicts discrete changes in performance on individual test questions at the admissions cutoff. This occurs because schools optimize their teaching time differently across skills. The most selective school might quickly cover basic skills

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

to spend more time on advanced material, while the less selective school might skip difficult skills entirely due to time constraints. These different teaching choices should create discontinuities in performance on specific questions that test different skills. Question-level discontinuities can exist even when overall test scores show no jumps for instance, if the test weights different skills in the same way that schools do when designing their curricula.

We now analyze how students perform on individual MCAS questions. This analysis faces several challenges. Since questions change yearly, we have fewer observations for each specific question. Individual question scores (0 or 1) provide noisier measurements than overall test scores. We're also limited to MCAS data, as question-level information isn't available for PSAT/SAT. Additionally, given the (relatively low) level of difficulty of the MCAS exam and the (relatively high) level of achievement of students who are around the BLS cutoff, many students near the cutoff either knew the material tested by some Grade 10 MCAS questions when they were in 6th grade, or they're capable of answering these questions correctly in 10th grade even without specific instruction. Despite these limitations, we find results that support the curriculum-matching model.

We begin by asking whether any individual questions show discrete changes in performance at the BLS cutoff. The answer is clearly yes. We ran 1,054 separate regression discontinuity analyses corresponding to one for each Grade 10 MCAS question, using a binary (0-1) outcome for each question instead of standardized scores. Figure 4 shows histograms of p-values for these estimated discontinuities, separated into English and Math sections. If there were no true discontinuities, these histograms should be roughly uniform across [0,1]. For example, only 5% of estimates should be significant at the 5% level. Instead, we find many more significant results: 16% of math questions and 12% of English questions show significant discontinuities at the 5% level. We also find notably fewer p-values above 0.5 than expected. For math exams, only 35% of questions have such p-values. These patterns demonstrate that performance discontinuities exist across many questions within each test and across both subjects.

The large number of estimates that are significant at the 5% level reflects that a number of questions for which the estimated upward or downward jump at the cutoff is quite large. As an illustration, Figure 5 provides RD plots for two questions from the same exam: questions 10 and 39 on the 2003 10th grade MCAS math exam. Question 10, shown on left, was answered correctly by 100% of (several dozen) students with entrance exam scores just below the BLS cutoff, but by only about 60% of students with scores just above the BLS cutoff. The identical set of students performed very differently on question 39. It was answered correctly by about 60% of students with scores just

below the BLS cutoff and by 80-90% of students with scores just above the cutoff.

We next examine test questions where performance gaps are largest to investigate the potential role of curriculum design, drawing on math curricula expertise developed by one of the authors in the course of writing math textbooks, Ellison (2010) and Ellison (2013). Figure 6 lists five Grade 10 MCAS Math questions where BLS students performed significantly worse (showing roughly 30 percentage point drops in performance). This includes the question corresponding to the RD plot on the left above, which was a probability problem about two spinners. The questions seem to fall into two distinct categories. The first group includes questions that favor recent middle school knowledge: the probability problem (2003 Q10), a proportional reasoning question about cylinder volume (2007 Q28) that's solvable by recognizing that tripling the radius leads to a nine-fold volume increase, and a problem (2007 Q30) that's more efficiently solved using pre-algebra trial-and-error rather than the more complex quadratic equation approach that high school students might attempt. The second group covers optional curriculum topics: a question about undefined slopes (2010 Q34), a concept that many schools might consider peripheral compared to core skills like using slope-intercept form, and a sequence problem (2006 Q1) that can be solved through various approaches.[20] The examples suggest BLS may have deliberately chosen not to emphasize these particular skills in their 9th and 10th grade curriculum, focusing instead on other mathematical concepts and approaches.

Figure 7 shows the five Grade 10 MCAS math questions where BLS students demonstrated the strongest performance advantage, with upward jumps of approximately 30 percentage points.[21] These questions show more diversity in content and difficulty. Two algebra questions (2014 Q11 and Q13) test fundamental skills that well-trained students should master, though some schools might avoid using numbers like 3 and 9 as exponents, and the negative numbers in Q13 could cause confusion. Another question asks students to write an equation for a "best fit" line on a scatterplot, a topic that wasn't traditionally part of high school curricula until its inclusion in the 2000 Massachusetts state standards. As of 2003, some schools serving students just below the BLS cutoff may not have updated their curricula to cover this material. The remaining two questions cover relatively basic concepts: rounding to the nearest whole number (a skill perhaps more commonly reinforced in science classes

[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 brute force approach that might occur to students who been studying systems of equations – assuming that the nth term is $an^2 + bn + c$ and using the values of the first three terms to find 3 equations in the unknowns a, b, and c – is much harder.

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

than math) and similar triangles (a topic that spans both middle and high school geometry). While these questions don't present as clear a pattern as the previous set, the results suggest that BLS students' stronger performance stems from more thorough training in advanced algebra and geometry concepts.

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

Prediction 4 presents another method for testing whether schools strategically allocate curriculum time based on student needs. We can subtract each student's "easy" test score from their "hard" test score to create a score difference variable. If schools are indeed tailoring their curriculum to their student population, we should observe a sudden increase in this score difference at the admissions cutoff.

The Grade 10 MCAS Math exam's design provides a valuable test case because it heavily features below-grade-level content. As detailed in Appendix A.2, the 2013 exam's alignment with Massachusetts's 2011 curriculum framework shows that 23 of 42 questions test middle school standards: 5 questions from grade six, 13 from grade seven, and 5 from grade eight. In contrast, the Grade 10 MCAS English exam operates at a more grade-appropriate level, combining challenging canonical high school literature (including translations of Greek and Latin works, Shakespeare, *Beowulf*, Austin, Cervantes, Conrad, Dickens, García Márquez, Kafka, and Shelley) with more accessible contemporary articles and texts.[22] The literary selections are approximately at a 10th-grade reading level, while the non-literary texts tend to be somewhat easier. Additional details about both exams' difficulty levels can be found in Appendix A.2.

We analyze four test score differences for each student, comparing their performance on harder versus easier exams. Specifically, we calculate: (1) highest SAT math score minus Grade 10 MCAS math score; (2) highest SAT critical reading score minus Grade 10 MCAS reading score; (3) Grade 11 PSAT math score minus Grade 10 MCAS math score; and (4) Grade 11 PSAT reading score minus Grade 10 MCAS English score. The PSAT comparisons offer the advantage of a smaller time gap, with only five months between tests. The SAT comparisons provide a larger difficulty gap between the tests. We normalize all scores to z-scores within each year's pool of BLS first-choice applicants for this calculation so that score differences are a natural measure.

[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*.

Estimates from this equation help differentiate between two competing models. The zero value-added model predicts no sudden changes in test score differences at the cutoff. In contrast, the curriculum-matching model (Prediction 4) suggests we should see upward jumps at the cutoffs, especially in the math score differences, as schools align their teaching to student ability levels.

Using the score difference as the dependent variable offers a potential side-benefit: it may reduce idiosyncratic noise. In equation (4), part of the noise term is the difference between actual and expected student ability, $E_{\theta|r_i}h(\theta) - h(r_i)$. This noise term cancels out in the score difference, provided that ability affects performance similarly on both exams. Additionally, differencing removes $E_{\theta|r_i}h(\theta)$ from the dependent variable, eliminating a potential source of nonlinearity that could bias RD estimation. This allows us to use wider bandwidths.[23] Examining score differences also has the potential to eliminate peer effects as an alternative explanation for discontinuities at the cutoff. To understand why, consider the curriculum-matching model, expanded to include peer effects. If a student's score on test k is

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

with the peer effect term $b\sum_{j\neq i} w_{ji}\theta_j$, not being test-dependent, then it will also cancel out when we difference leaving curriculum match quality as the only remaining explanation for any observed discontinuity at the cutoff.

Figure 8 displays our analysis of test score differences. We group BLS applicants who scored within 20 percentiles of the cutoff into one-percentile bins. Each dot represents the bin's average difference between students' standardized PSAT/SAT scores and their MCAS scores. Local linear regression estimates trace the conditional means. The top panels show comparisons of math tests. On the left, PSAT math scores show a substantial upward jump (approximately 0.16σ) at the BLS cutoff. The right panel reveals a smaller but clear upward jump in SAT math scores. The bottom panels compare reading tests, where there are discontinuities of roughly 0.09σ for each outcome.

Table 3 provides formal statistical analysis of the discontinuities shown in the figure, using local linear estimates of equation (4) with test score differences as dependent variables. Both math comparisons (PSAT-MCAS and SAT-MCAS) show significant upward jumps of approximately 0.16σ and

[23] Under the null hypothesis of no curriculum-matching effects, the relationship between score difference and entrance exam performance should have the same second derivative on both sides of the cutoff. This property reduces concerns about bias when using wider analysis windows. Furthermore, we wouldn't expect the conditional expectation, $\mathbb{E}(h_k(\theta)|r)$, to show strong nonlinearity near the cutoff, since students at the cutoff score around the 70th percentile on MCAS - well within the middle range of the distribution rather than in its tails.

0.08σ at the cutoff, respectively. Both English comparisons have jumps of 0.08, though these are less precisely estimated than the math comparisons. These findings support the curriculum-matching model over the zero value-added model in explaining how BLS affects student performance across tests of varying difficulty.

5 Advanced Placement and College Outcomes

We next turn to measures of more advanced educational outcomes, targeted towards the high-achieving students at BLS. The Advanced Placement (AP) program offers college-level material to high school students, with tests scored from 1-5. Many colleges grant credit for scores of 3 or higher, which we define as passing. Following Abdulkadiroğlu, Angrist, and Pathak (2014), we focus on popular and very popular AP subjects. This typically includes core subjects like math, science, English, and history, while excluding arts, music, and foreign languages.[24] BLS's enrollment of high-achieving students creates sufficient demand to offer an extensive AP curriculum. Table 1 shows striking differences in AP participation: while the average Boston 6th grader eventually takes 0.7 AP tests, students near the BLS admissions cutoff take an average of 2.5 tests.

BLS students, while representing only about 10% of Boston Public Schools' high school enrollment, dominate AP test-taking and achievement. In 2022, they took 2,341 AP tests with an 84.2% pass rate (scores of 3 or higher). This accounts for 38% of all AP tests taken in Boston Public Schools and 58% of passing scores (1,971 out of 3,379). Their dominance is even more pronounced in advanced senior-year courses: BLS students earned 82% of all passing scores on AP Calculus BC and 84% on AP English Literature among Boston Public Schools students.[25] These AP participation patterns align with our curriculum-matching model: BLS's concentration of high-achieving students enables the school to effectively offer and teach these advanced subjects.

RD analyses indicate that one reason for this dominance is a causal effect: BLS admission significantly improves AP test-taking and performance, when measured via all AP exams or very popular AP exams.[26] First, as shown in Panel A of Table 4, a BLS offer increases AP exam participation by

[24] Very popular AP subjects are US History, Biology, English Language and Composition, English Literature, US Government and Policies, and Calculus AB. Popular tests include very popular AP tests as well as Calculus BC, Statistics, Chemistry, Physics B/C, European History, Microeconomics, and Macroeconomics.

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

[26] Given the discreteness of AP scores, 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.

0.6 tests overall and by 0.3 tests for very popular AP subjects. These effects are substantial relative to the control group's mean participation. Second, Panel B demonstrates that a BLS offer increases the number of AP tests passed by 0.39 and the number of very popular AP tests passed by 0.15.[27] The estimates on test-taking and performance for popular AP exams are positive, but are not statistically significant. The overall picture of estimates on AP participation and performance support the curriculum-matching model over the zero-value added model: concentrating high-ability students in one school enables the offering of advanced curricula like AP courses.

Our final outcome is college attendance and completion. As noted earlier, college completion is a challenge for BPS students. Even among high-achieving students scoring near the BLS admissions cutoff, Table 1 shows that about one-third of college enrollees fail to graduate. BLS has strong connections to elite colleges. For instance, Bernhard (2013) notes that BLS was among the nation's top seven feeder schools to Harvard in 2013, sending 13 students. However, since selective colleges often cap admissions from individual high schools, it is unclear whether BLS would boost elite college admission chances for students who barely met BLS's entrance requirements, and college dropout is quantitatively more important than failure to enroll in college for students scoring near the BLS cutoff.

Table 5 shows that BLS admissions has positive effects on several college outcomes, but none are statistically precise enough to rule out a chance finding. For example, at the BLS cutoff, students who are admitted show a 3.1 percentage point increase in overall college completion and 5.5 percentage point increase in four-year college graduation rates, but the standard errors of both estimates are about 4.5.[28] Part of the difficulty in identifying discontinuities at the BLS admissions cutoff is the need to distinguish them from increases in the slope of the admissions-graduation relationship at the admissions cutoff. The latter are significant in two of the regressions, and this could reflect that students who are better-matched to the BLS curriculum benefit even more from admission, but we are not treating any such benefits as estimated causal effects.

[27] 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 popular AP tests compared to standard errors of 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.

[28] Abdulkadiroğlu, Angrist, and Pathak (2014) found positive effects of BLS admission on college attendance for 7th and 9th grade applicants (6.2 percentage points overall and 10 percentage points for four-year colleges, the latter significant at 5%). Our analysis shows smaller, non-significant effects though our confidence intervals include their estimates.

6 Conclusion

This paper presents two simple models of school system and curriculum design. In a system where students have different learning styles, aggregate achievement is maximized by matching students to schools based on learning styles, with schools then tailoring their curriculum to their students' needs. Our analysis shows that regression discontinuity estimates at selective school admission cutoffs cannot distinguish between two different explanations: either (1) the selective school adds no value, or (2) both student allocation and curriculum are optimally designed to maximize total achievement.

Our findings reveal how the curriculum-matching model makes distinct predictions that help differentiate between explanations. Under curriculum matching, we expect: (1) the relationship between entrance exam scores and later performance changes slope at admission cutoffs, as students move from mismatched to well-matched curricula; (2) students prefer the selective school before knowing their scores, since this preference only matters if they qualify; (3) performance on individual test questions shows discontinuities, reflecting different time allocations to topics across schools; and (4) near the admission cutoff, students at selective schools perform better on challenging tests, as these schools emphasize advanced skills.

Testing the predictions of the curriculum-matching model involves several empirical challenges. Detecting changes in performance slopes at admission cutoffs requires substantially more data than identifying simple changes in levels. Analyzing individual test questions introduces additional complexity, as these analyses rely on binary outcomes from single-year data (since test questions change annually). Despite these methodological hurdles, our findings support several predictions of the curriculum-matching model. The most compelling evidence emerges from what we expected to be the most revealing test: students admitted to Boston Latin School demonstrate better performance when measuring the gap between their PSAT and Grade 10 MCAS results and the gap between their SAT scores and Grade 10 MCAS results for both Math and English.

The curriculum-matching model initially compared schools that valued all skills equally but specialized due to students' different learning rates. However, the value of mastering specific skills likely varies among students. For college-bound students, exposure to college-level texts and STEM-preparatory math may be crucial. For others, mastering practical skills tested on graduation exams may be more valuable for daily life.

This analysis highlights the importance of aligning test score measures with educational goals when evaluating interventions. Tests vary in both content and how they translate mastery into scores.

Schools teach a variety of skills that benefit students both in the short term and over their lifetimes. Failing to recognize these distinctions can result in misleading conclusions about the value of selective schools and the impact 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.

Aitken, C., G. Gray-Lobe, M. Joshi, M. Kremer, J. de Laat, and W. Wong (2025): "Hard to Read: The Impact of Advanced Reading Assignments on Language and Literacy Outcomes," Working paper, University of Chicago.

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

Bau, N. (2022): "Estimating an Equilibrium Model of Horizontal Competition in Education," *Journal of Political Economy*, 130(7), 1717–1764.

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

——— (2025): "Targeted Acceration in Middle School Math: Impacts on College Entry, Degreee Completion, and STEM," Working Paper.

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).

Cohodes, S. R. (2016): "Teaching to the Student: Charter School Effectiveness in Spite of Perverse Incentives," *Education Finance and Policy*, 11(1), 1–42.

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.

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.

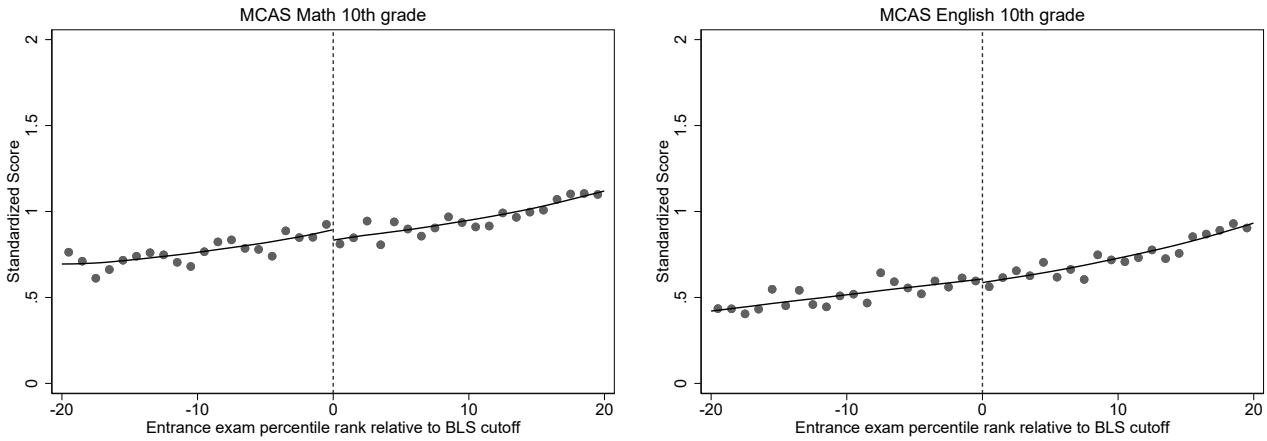


Figure 1: Performance on 10th grade MCAS vs. admissions test score.

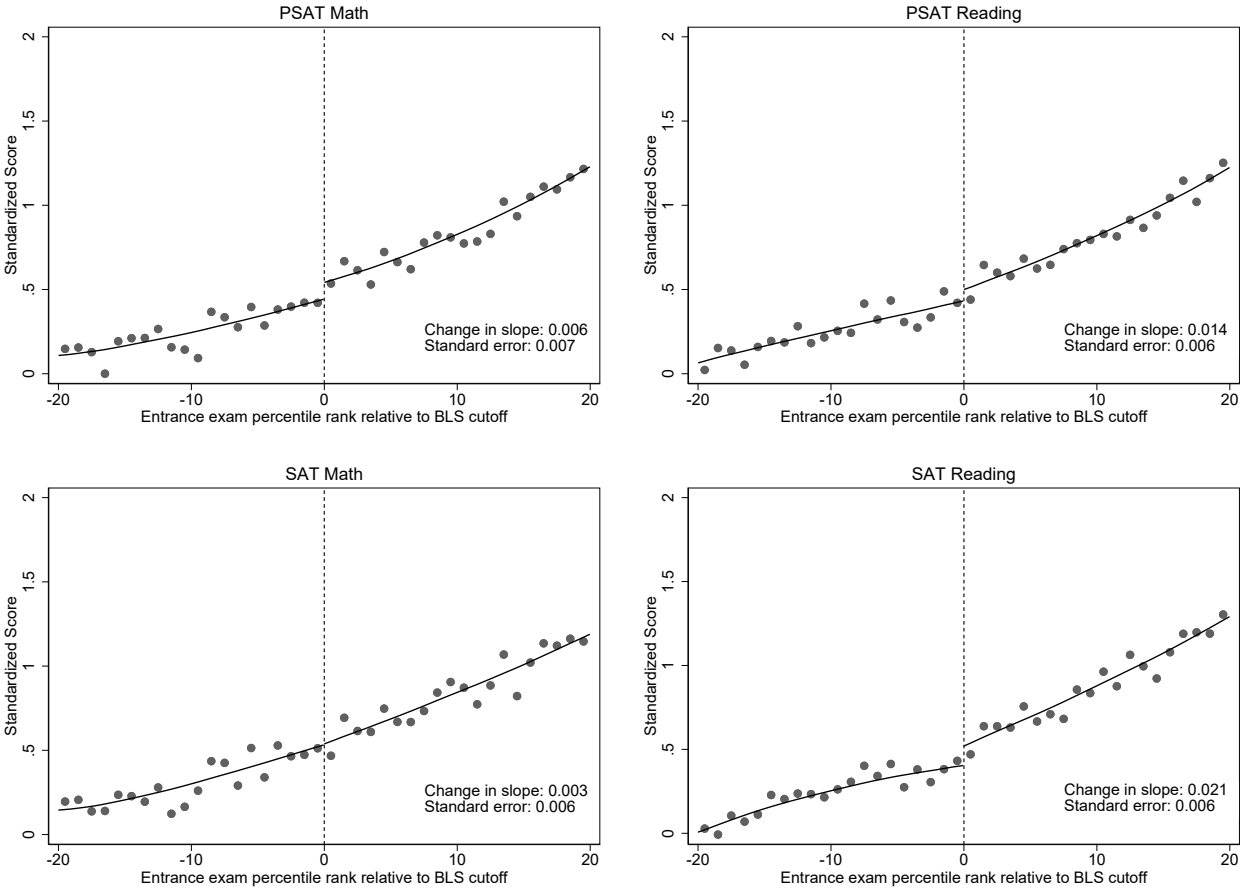


Figure 2: Slopes of the achievement vs. entrance score relationship around the admissions cutoff

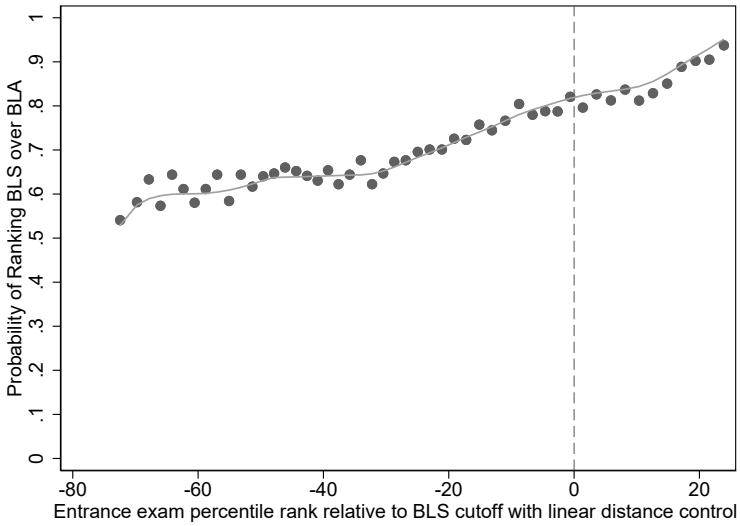


Figure 3: Fraction of applicants for grade 7 preferring Boston Latin School (BLS) to Boston Latin Academy (BLA) by admissions exam percentile relative to the BLS cutoff

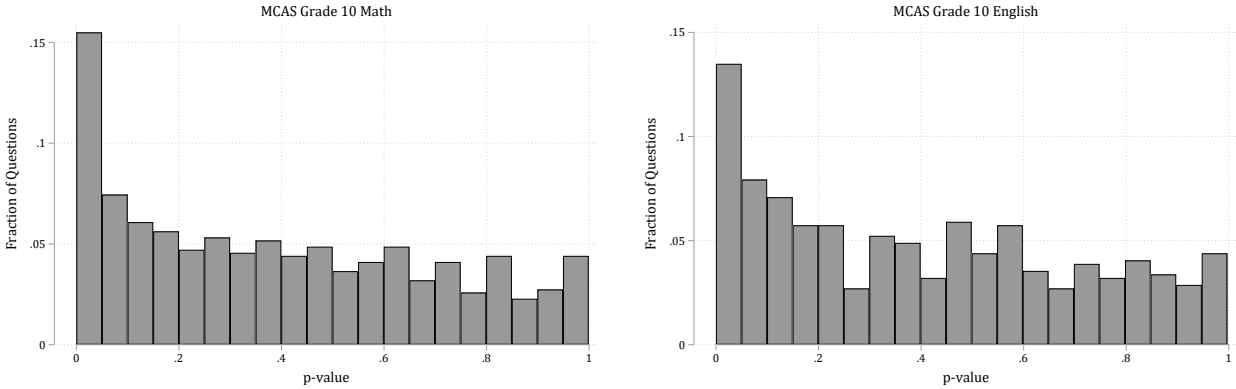


Figure 4: P-values of jumps at the Boston Latin School cutoff on individual Grade 10 MCAS questions.

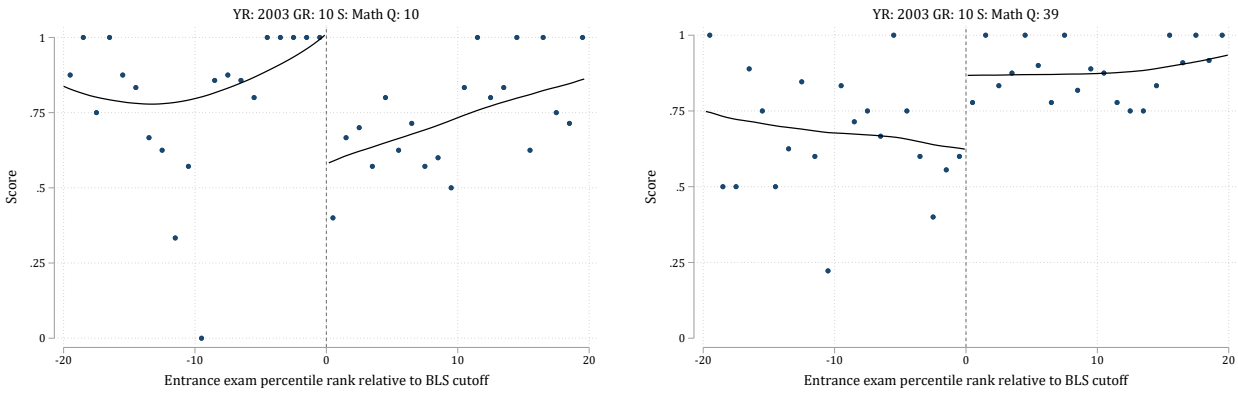
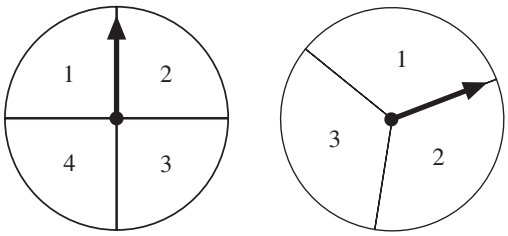


Figure 5: Some MCAS math questions with large apparent jumps at the cutoff: 2003 grade 10 MCAS questions 10 and 39

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?

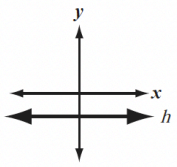
A. 4 inches
B. 6 inches
C. 9 inches
D. 18 inches

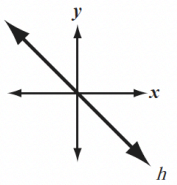
1 The first five terms in a quadratic sequence are shown below.

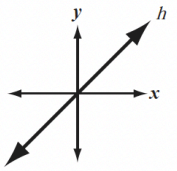
6, 9, 14, 21, 30, . . .

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.


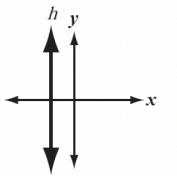
D.


Figure 6: Sample MCAS questions with most statistically significant 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?

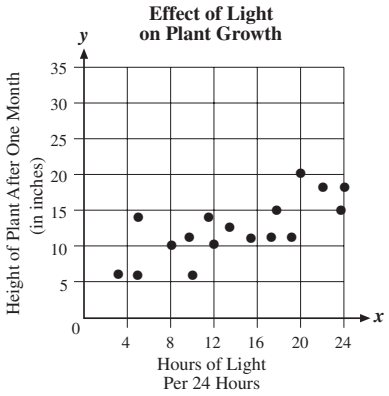
$$-2(x - 5)$$

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

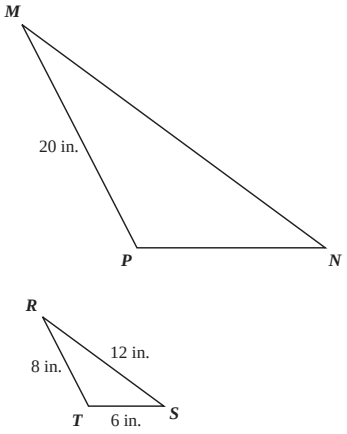
39 Jenny studied the effect of light on plant growth. She graphed a scatterplot to represent her data.



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 7: Sample MCAS questions with most statistically significant upward jumps at the Boston Latin School cutoff

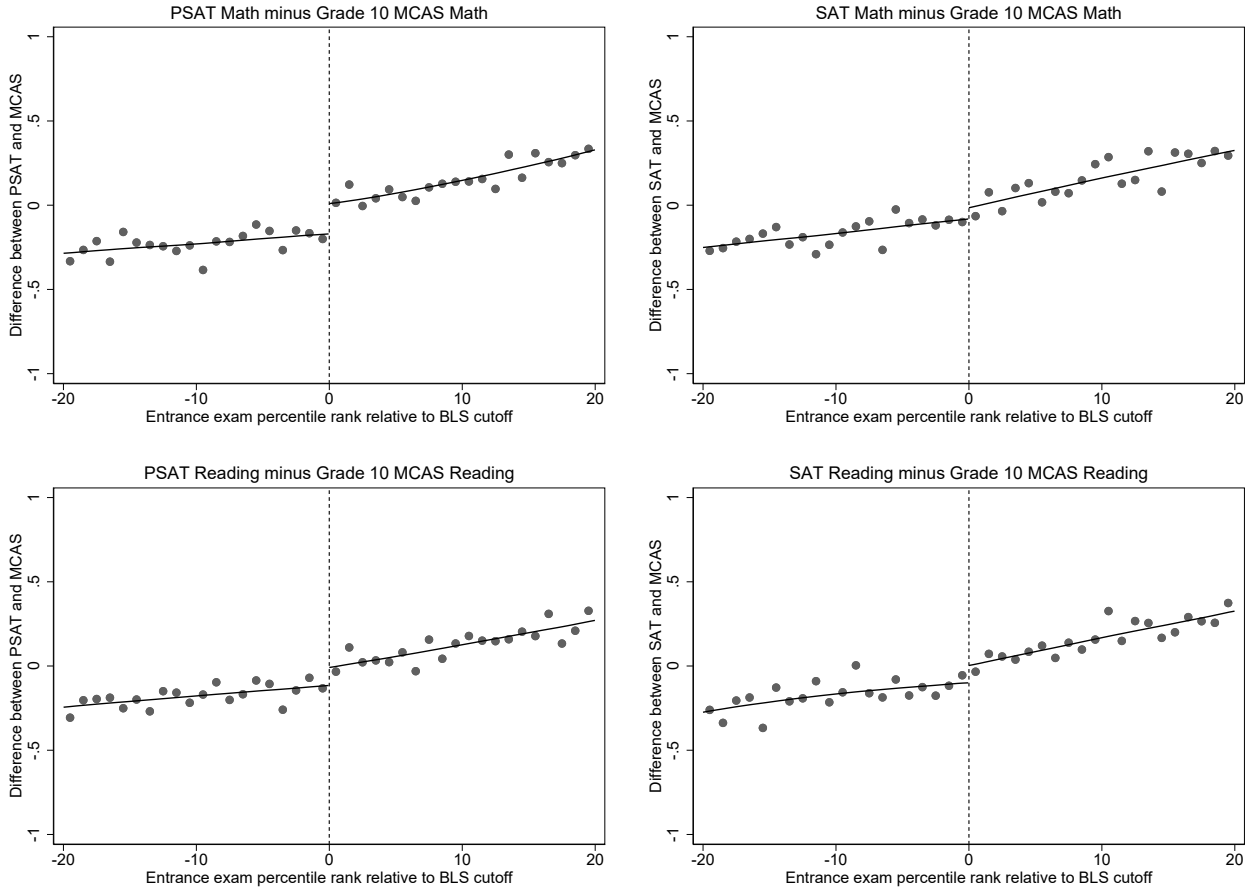


Figure 8: Differences between scores on more and less difficult tests: PSAT/SAT score minus 10th grade MCAS score at the Boston Latin School cutoff

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	0.44	0.33	0.28	0.18
Hispanic	0.34	0.24	0.23	0.18
White	0.13	0.21	0.23	0.31
Asian	0.08	0.21	0.25	0.33
	76,496	19,862	13,572	5,492
		B. MCAS test outcomes		
Grade 6 Math	-0.43	0.26	0.40	0.95
Grade 6 English	-0.51	0.11	0.23	0.66
Grade 10 Math	-0.29	0.39	0.50	0.86
Grade 10 English	-0.38	0.20	0.30	0.62
		C. Pre-college test outcomes		
PSAT Math	42	47	49	53
PSAT English	39	43	45	49
SAT Math	472	534	554	600
SAT English	449	507	525	569
Number of APs Taken	0.69	1.53	1.79	2.51
w/ AP Score ≥ 3	0.31	0.88	1.12	1.69
w/ Score ≥ 3 on Popular APs	0.21	0.61	0.76	1.14
w/ Score ≥ 3 on Very Popular APs	0.14	0.38	0.48	0.72
		D. Post-secondary outcomes (among NSC-queried)		
Attend Any College	0.60	0.81	0.83	0.88
Attend 4 year College	0.42	0.68	0.72	0.83
Persist Any College	0.55	0.78	0.80	0.87
Persist 4 year College	0.39	0.66	0.70	0.81
Graduate Any College	0.22	0.43	0.48	0.59
Graduate 4 year College	0.18	0.40	0.44	0.57

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. We defined two exam groupings: Popular exams, those with at least 500 test-takers as identified in Abdulkadiroğlu, Angrist, and Pathak (2014), including U.S. History, Biology, Chemistry, Microeconomics, Macroeconomics, English Language, English Literature, European History, U.S. Government, Calculus AB, Calculus BC, Physics B, Physics C: Mechanics, Physics C: Electricity and Magnetism, and Statistics; and a narrower Very Popular subset, those with at least 1,000 test-takers, consisting of all exams in Popular exams and U.S. History, Biology, English Language, English Literature, U.S. Government, and Calculus AB. Number of APs Taken "w/ Score ≥ 3" counts the number of AP tests with a 3 or higher. Number of APs Taken with "w/ Score ≥ 3 counts on Popular APs" counts the number of Popular AP tests with a 3 or higher. Number of APs Taken with "w/ Score ≥ 3 counts on Very Popular APs" counts the number of Very Popular AP tests with a 3 or higher. 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 ($\boldsymbol{\rho}$)	-0.034	0.074	0.025	0.000	0.047	0.122**
	(0.029)	(0.049)	(0.047)	(0.033)	(0.045)	(0.048)
Change in slope at cutoff ($\boldsymbol{\gamma}$)	-0.004	0.005	0.003	0.005	0.014**	0.021***
	(0.004)	(0.006)	(0.006)	(0.004)	(0.006)	(0.006)
Control mean	0.77	0.25	0.31	0.51	0.26	0.25
Observations	3349	3200	3393	3933	3467	3120
Bandwidth	13.83	14.79	16.47	16.20	16.11	15.05

Notes: This table reports estimates of changes in levels and the slope of achievement outcomes at the BLS admissions cutoff. Control mean reports the outcome means for the [-20, 0] window. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 ($\boldsymbol{\rho}$)	0.159***	0.072**	0.087*	0.089*
	(0.041)	(0.035)	(0.046)	(0.048)
Control mean	-0.23	-0.16	-0.17	-0.16
Observations	3864	4120	4295	4123
Bandwidth	18.23	29.33	26.31	20.79

Notes: This table displays the changes in differences in academic achievement measured as the difference between a student's highest score on the math/critical reading SAT and their score on the Grade 10 Math/English MCAS and the difference between a student's score on the math/reading PSAT and the Grade 10 Math/English MCAS. Control mean reports the outcome means for non offers within [-20, 20] window. 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 (1)	Popular AP Exams (2)	Very Popular AP Exams (3)
		A. Test Taking	
Change in level at cutoff (ρ)	0.617***	0.141	0.266***
	(0.161)	(0.124)	(0.084)
Change in slope at cutoff (γ)	-0.019	-0.007	-0.001
	(0.025)	(0.019)	(0.013)
Control mean	1.75	1.33	0.84
Observations	3140	3168	3127
Bandwidth	12.91	13.01	12.87
		B. Number of AP Tests Passed	
Change in level at cutoff (ρ)	0.387***	0.027	0.152**
	(0.147)	(0.115)	(0.074)
Change in slope at cutoff (γ)	0.008	0.001	0.004
	(0.023)	(0.020)	(0.012)
Control mean	0.93	0.69	0.43
Observations	3063	2844	2961
Bandwidth	12.59	11.73	12.18

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 Popular and Very Popular AP tests. Very Popular AP tests are defined as U.S. History, Biology, English Language and Composition, English Literature and Composition, U.S. Government and Politics, and Calculus AB. Popular AP tests are defined as those including all Very Popular AP tests as well as 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. 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)	0.005	0.001	0.031
	(0.023)	(0.025)	(0.045)
Change in slope at cutoff (γ)	-0.002	-0.002	0.008*
	(0.003)	(0.003)	(0.004)
Control mean	0.86	0.84	0.53
Observations	3658	3305	2298
Bandwidth	17.26	18.80	19.86
		B. 4 Year College	
Change in level at cutoff (ρ)	0.031	0.037	0.055
	(0.026)	(0.030)	(0.046)
Change in slope at cutoff (γ)	0.004*	0.005	0.008
	(0.002)	(0.003)	(0.005)
Control mean	0.78	0.77	0.50
Observations	4276	3420	2156
Bandwidth	20.58	19.32	18.81

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. Control mean reports the outcome means for non-offers within [-20, 20] window. 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 Grade 10 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 Grade 10 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.[29]

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

[29] 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.[30] Table A1 list the Lexile indexes we could find for books that had excerpts used as reading passages on 2004-2013 Grade 10 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.[31] 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

[30] See Hiebert (2009).
[31] 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)	-0.004	0.007	0.026	-0.016	0.014	
	(0.019)	(0.023)	(0.024)	(0.023)	(0.024)	
Control mean	0.89	0.82	0.79	0.81	0.78	
Observations	4227	4533	4812	4651	4933	
Bandwidth	15.49	17.55	18.87	18.03	19.29	
			B. English			
Change in level at cutoff (ρ)	-0.011	0.026	0.007	-0.014	0.017	
	(0.020)	(0.024)	(0.023)	(0.024)	(0.024)	
Control mean	0.89	0.79	0.82	0.81	0.78	
Observations	3872	4812	4533	4486	4662	
Bandwidth	14.25	18.87	17.55	17.38	18.35	
			C. Other Outcomes			
Change in level at cutoff (ρ)						-0.013
						(0.011)
Control mean						0.86
Observations						4721
Bandwidth						17.33

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

The analysis draws on a comprehensive set of administrative records from Boston Public Schools (BPS) and the Massachusetts Department of Elementary and Secondary Education (DESE), supplemented by standardized testing agencies and the National Student Clearinghouse (NSC). We begin with BPS data on exam school applicants from 1995 to 2017, restricting the sample to those applying to the 7th grade between 1999 and 2015. These records include student demographics, school preferences, and composite admission scores, and are cleaned to exclude ineligible or inconsistent cases. We use crosswalks connecting BPS student identifiers to state-assigned student identification numbers (SASIDs) to link datasets. Enrollment trends are compiled from both BPS (1995–2012) and DESE's Student Information Management System (SIMS) (2008–2019). Academic achievement is measured using the Massachusetts Comprehensive Assessment System (MCAS) scores in Math and English Language Arts (ELA) for grades 3–10 (2002–2019), including both scaled scores and Grade 10 item-level responses. We obtained Scholastic Assessment Test (SAT) and Preliminary SAT (PSAT) scores from BPS and DESE. These scores were standardized by subject and year, utilizing College Board concordance tables to accommodate changes in test formats. Advanced Placement (AP) data from 2005 to 2020 were sourced from BPS for the earlier years and from DESE in subsequent years. Our NSC post-secondary outcomes, spanning 2007 to 2020, include information on college enrollment, persistence, and graduation rates, which were linked through SASIDs with appropriate timing adjustments for each cohort. Detailed preparations for each dataset are elaborated in the subsections below.

Exam School Applicants Data

We use applicant-level data from BPS covering exam school applications from 1995 to 2017 school years. The dataset includes a record for each applicant, detailing their application ID, SASID, name, gender, race, date of birth, application year, grade level at application, ranked preferences for the three exam schools, and their composite admission score. To ensure analytic validity, we restricted the sample to students who met the criteria for valid exam school applicants. We begin by dropping students who applied to grade 10 and those outside the application year window of 1999-2015. We remove duplicate observations within each school year and application grade. We exclude students applying from private schools, those who did not apply to any exam school, and those without a rank for any exam school. Additionally, we drop students who received offers from schools they did not rank, as well as students who were not offered admission despite having scores above the average admitted student. The analysis sample includes exam school applicants applying to the 7th grade for school years 1999 through 2015. These years correspond to those for which we have available outcomes.

Crosswalk Data

To merge BPS administrative records with DESE outcome files, we constructed a crosswalk linking key identifiers across datasets. For the student identifier crosswalk linking BPS IDs to state-level SASIDs, we assembled a comprehensive panel of BPS assignment and enrollment records spanning 1997 to 2019.

For the 1997–2005 cohorts, student records were extracted from BPS assignment files and merged with enrollment files using an internal crosswalk file to retrieve names and dates of birth. For 2006–2013, we processed over two dozen rounds of assignment files from BPS archives, standardizing student names and dates of birth. For 2014–2019, data were drawn from BPS assignment megafiles. Across all years, we standardized formatting, resolved naming inconsistencies, and removed duplicate records to construct a unified student-level file. These BPS records were matched to DESE SIMS files from 2001 to 2019, using combinations of student number, name, and date of birth. When available, we preferred SIMS-based matches; otherwise, we relied on earlier internally constructed crosswalks. In cases of conflicting SASID matches, we prioritized matches with the smallest date-of-birth discrepancies.

Enrollment Data

The enrollment data combine records from BPS and DESE. For 1995–2012, we use cleaned BPS enrollment files, while for 2008–2019, we rely on DESE's SIMS. Due to gaps in earlier state records, BPS data are critical for capturing the full enrollment history of exam school applicants. The overlapping years (2009–2012) provide a check on consistency across sources. To construct the BPS portion, we combined multiple raw files, standardized year and grade formats, cleaned key demographic and program participation fields, and calculated cumulative years spent in exam schools. For DESE data, we processed fall and end-of-year SIMS submissions, harmonized variable definitions across years, and merged observations to retain the most complete student-school-year record. The final merged file spans 1995 to 2019 and includes variables such as school attended, special education status, subsidized lunch eligibility, and English proficiency. When students appear in multiple schools in a year, we assign them to the school with the longest enrollment. This cleaned enrollment dataset is then merged with the application file to identify BPS students applying to exam schools between 1999 and 2015.

MCAS Data

The MCAS dataset, provided by DESE, covers the years 2002 to 2019 and includes both raw and scaled scores in Math and ELA for students in grades 3 through 10. It also contains item-level data indicating whether each question was answered correctly, incorrectly, or left blank. From this dataset, we construct two files. The first file includes raw and scaled scores for MCAS Math and ELA across all grades and students. This file is merged with the application and enrollment datasets to create our main analysis file. Prior to merging, we address a small number of cases where students have raw scores but no corresponding scaled scores by imputing scaled scores using the observed raw-to-scaled score relationship among students who took the same test. Scaled scores are then standardized to have a mean of zero and a standard deviation of one within each subject-grade-year, based on all Massachusetts students with MCAS scores. The second file contains question-level data for every multiple-choice item on the Grade 10 MCAS Math and ELA exams for each year, excluding 2016 due to missing item-level data. For each question, we compute the number of students who answered it correctly. Importantly, the Grade 10 MCAS exams were not affected by the PARCC (Partnership for Assessment of Readiness for College and Careers) changes in 2015 and 2016.

SAT Data

The SAT dataset is compiled from records provided by both BPS and DESE. For school years 2005 and 2006, SAT scores are sourced from BPS and include student identifiers (BPS IDs) along with raw scores for the Math, Reading, and Writing sections. These records are linked to SASIDs using a crosswalk file. From 2007 to 2020, SAT scores are sourced from DESE and already include SASIDs. In both cases, BPS and DESE receive SAT data for exam school applicants directly from the College Board. If a student took the SAT multiple times, we retain the highest total score (Math + Reading) across all test attempts. To account for the 2016 redesign of the SAT, we convert pre-redesign scores using official concordance tables provided by the College Board.[32] We append the BPS and DESE records to construct a consolidated SAT dataset. Prior to merging with the analysis file, we correct data entry errors by multiplying any score recorded below 200 by 10. After the merge, raw SAT scores are standardized to have a mean of zero and a standard deviation of one within each subject-year, based on exam school applicants in our sample who took the test in that year. Additionally, we set SAT outcomes to missing for the 2015 application cohort, as the available SAT data through 2020 do not provide complete coverage for that group.

PSAT Data

All PSAT data were provided by BPS. In the earlier years (2004–2005), students are identified using BPS-specific student IDs, which we link to SASIDs via a crosswalk. For the years 2006–2019, the data already include SASIDs, eliminating the need for this step. While earlier datasets include scores for all BPS students, the later years contain scores only for exam school applicants. Since our analysis focuses exclusively on exam school applicants, we restrict the sample accordingly in all years. If a student took the PSAT more than once, we prioritize their grade 11 score. The PSAT scoring scale was redesigned in 2016. To ensure comparability across years, we convert pre-redesign scores using the concordance tables provided by BPS and the College Board.[33] Following the redesign, the Reading and Writing sections were combined into a single Evidence-Based Reading and Writing (ERW) score. Once scores are converted, the ERW score is divided by two to recover separate Reading and Writing components. The PSAT data are then appended across all years, and records with valid SASIDs are merged with the application, enrollment, MCAS, and SAT files. Following the merge, raw PSAT scores are standardized to have a mean of zero and a standard deviation of one within each subject-year, using only exam school applicants in our analytic sample tested in that year. Additionally, we set PSAT outcomes to missing for the 1999 application cohort, as the available PSAT data from 2004–2005 do not provide full coverage for that group.

AP Data

Our AP data span the years 2005 to 2020 and are drawn from two primary sources. For 2005 and 2006, student-level AP scores were provided by BPS, which identified students using internal BPS

[32] See Concordance Tables from the College Board.

[33] Available at: `https://studylib.net/doc/18228722/psat-s̄nmsqt-s̄understanding-s̄scores-s̄2015`

IDs. These were linked to SASIDs using a crosswalk file. From 2007 onward, AP records were sourced from DESE and already included SASIDs. After appending all years, we constructed a unified dataset of AP outcomes keyed by SASID. For each student, we calculated the number of AP exams taken per year and total scores. To analyze patterns in subject choices, we defined two exam groupings: Popular exams—those with at least 500 test-takers as identified in Abdulkadiroğlu, Angrist, and Pathak (2014)—including U.S. History, Biology, Chemistry, Microeconomics, Macroeconomics, English Language, English Literature, European History, U.S. Government, Calculus AB, Calculus BC, Physics B, Physics C: Mechanics, Physics C: Electricity and Magnetism, and Statistics; and a narrower Very Popular subset—those with at least 1,000 test-takers—consisting of all exams in Popular exams and U.S. History, Biology, English Language, English Literature, U.S. Government, and Calculus AB. The resulting dataset was merged with application, enrollment, MCAS, SAT, and PSAT records. Students without any AP data were assigned zeros on all AP outcome variables.

NSC Data

Data on college enrollment come from the NSC, which provides post-secondary enrollment records based on submissions from DESE. DESE submits student names and birthdates for in-state high school students, with NSC conducting annual searches for graduates and biennial searches for non-graduates. We use NSC data from the 2007–08 through 2019–20 school years to construct college attendance, persistence, and graduation indicators for Massachusetts students. After importing and cleaning the raw NSC file, we standardized variables, and modified college codes. We then generated indicators for attendance (at least one semester), persistence (four or more semesters), and graduation, separately for 2-year and 4-year colleges. These outcomes were collapsed to one record per student, and merged with other datasets (application, enrollment, MCAS, SAT, PSAT, and AP) using SASIDs. To ensure alignment with application cohorts, we set attendance outcomes to missing for years beyond 2012, persistence beyond 2010, and graduation beyond 2006, reflecting typical timelines for post-secondary progression.

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.