Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

High-School Exit Examinations and the Schooling Decisions of Teenagers: Evidence From Regression-Discontinuity Approaches

John P. Papay , Richard J. Murnane & John B. Willett

To cite this article: John P. Papay , Richard J. Murnane & John B. Willett (2014) High-School Exit Examinations and the Schooling Decisions of Teenagers: Evidence From Regression-Discontinuity Approaches, Journal of Research on Educational Effectiveness, 7:1, 1-27, DOI: 10.1080/19345747.2013.819398

To link to this article: https://doi.org/10.1080/19345747.2013.819398

Published online: 10 Jan 2014.

Submit your article to this journal

Article views: 323

View related articles

View Crossmark data

Citing articles: 7 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uree20

Journal of Research on Educational Effectiveness, 7: 1–27, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2013.819398

Routledge
Taylor & Francis Group

INTERVENTION, EVALUATION, AND POLICY STUDIES

High-School Exit Examinations and the Schooling Decisions of Teenagers: Evidence From Regression-Discontinuity Approaches

John P. Papay
Brown University, Providence, Rhode Island, USA

Richard J. Murnane
Harvard University, Cambridge, Massachusetts, USA

John B. Willett
Harvard Graduate School of Education, Cambridge, Massachusetts, USA

Abstract: We examine whether barely failing one or more state-mandated high school exit examinations in Massachusetts affects the probability that students enroll in college. We extend the exit examination literature in two ways. First, we explore longer term effects of failing these tests. We find that barely failing an exit examination, for students on the margin of passing, reduces the probability of college attendance several years after the test. Second, we explore potential interactions that arise because students must pass exit examinations in both mathematics and English language arts in order to graduate from high school. We adopt a variety of regression-discontinuity approaches to address situations where multiple variables assign individuals to a range of treatments; some of these approaches enable us to examine whether the effect of barely failing one examination depends on student performance on the other. We document the range of causal effects estimated by each approach. We argue that each approach presents opportunities and limitations for making causal inferences in such situations and that the choice of approach should match the question of interest.

Keywords: Exit examinations, high-stakes tests, regression-discontinuity

For much of the 20th century, students in the United States graduated from high school, entered college, and completed a 4-year college degree at much greater rates than their parents. These rising educational attainments contributed to improved living standards for several generations of Americans (Goldin & Katz, 2008). This is one reason that increasing the educational attainments of today's children has been a central focus of education policy over the past decade. In particular, boosting college enrollment is seen as a critical step to ensuring that students gain the skills necessary to succeed in the 21st-century American economy.

Why do so many American teenagers drop out of the educational pipeline before they enter college? There are myriad explanations: poor information about labor markets, credit

Address correspondence to John P. Papay, Education Department, Brown University, Providence, RI 02912, USA. E-mail: john_papay@brown.edu

constraints that prevent borrowing against future income, rising costs of college attendance, and perceived high nonpecuniary costs of succeeding in school. Here, we focus on one element of nonpecuniary costs that may be important for academically struggling students: exit examinations that high school students must pass to obtain a diploma. Today, more than three fourths of American public school students face high school exit examination requirements (Center on Education Policy, 2010). Typically, students take these tests early in their high school career. Although students have multiple opportunities to retake them, failing one or more exit examinations early in a high school career may increase the cost of graduation because students must prepare for, and spend time retaking, the tests.

The literature on high school exit examinations has focused largely on high school graduation as the key outcome. In other words, researchers have examined whether barely failing an exit examination increases dropout rates. However, barely failing an exit examination may have longer term consequences for students, for example, by reducing their probability of attending college. To the extent that failing an exit examination induces some students to drop out of high school, the costs of postsecondary access are increased as students generally must take and pass the GED examinations to enroll in college. The high-stakes tests may also pose additional psychic barriers to continuing in school, as students who fail an examination may then lose confidence in their academic abilities. As a result, failing an exit examination may affect not only the probability that students graduate from high school but also their postsecondary educational attainments.

For researchers, analyzing the effect of high school exit examination performance on educational attainments is both facilitated and complicated by the nature of the requirements themselves. Students either pass or fail the examinations based on their performance relative to exogenously assigned cut scores. Because of this, and providing the underlying assumptions are satisfied, researchers can use a regression-discontinuity approach to obtain unbiased estimates of the causal effect of examination performance on educational attainments, for students in the immediate vicinity of the cutoff. In particular, such approaches can help policymakers understand whether barely passing an examination on the first try, as opposed to barely failing it, affects the subsequent educational attainments of students on the margin of passing. Any such effects represent the consequences of failing the test for students scoring at the cutoff. We can think of these as unintended consequences of the examination policies because students whose examination scores fall just on either side of the cutoff are essentially equally skilled. Consequently, policymakers should not want to treat equally proficient students differently simply based on which side of the cutoff their test scores fall. In this article, we build on our prior work, looking beyond the effects on high school graduation to examine the longer lasting effects of examination performance on college-going, an especially important outcome given the very large differential between the average labor-market earnings of college graduates and terminal high school graduates.

The case of high school exit examinations also proves interesting methodologically because it is one of a growing set of situations in which individuals are assigned to a range of treatments based on their positions relative to exogenously assigned cutoffs on multiple forcing variables (also called "rating scores" or "assignment variables"). Here, for instance, students may have to pass several examinations (e.g., in mathematics and English language arts). In these cases, researchers and policymakers may be interested in a wide range of causal effects, based on comparisons at the cutoffs by one rating score, the other, or some conditional combination of them. There is growing interest in the more nuanced analyses that such designs can support.

Reardon and Robinson (2012) presented a range of options that researchers can use to examine such situations, and they described the theoretical trade-offs involved in adopting

these approaches. In this article, we build on this framework and apply these approaches to examine a substantive question using rich data. In addition, we explain and implement an approach that we developed (Papay, Willett, & Murnane, 2011) to examine such situations more flexibly. We assess the strengths and weaknesses of these approaches empirically and draw substantive conclusions about the causal effect of barely passing exit examinations on subsequent college attendance. We begin with a brief discussion of how high school exit-examination performance can affect student educational outcomes. We then describe our analytic approach and the Massachusetts data we draw on. We present our results, describe the threats to the validity of our main inferences, and conclude with methodological lessons and policy implications of this work.

THE EFFECTS OF FAILING A HIGH SCHOOL EXIT EXAMINATION

In a simple economic model, students will continue to invest in schooling if their assessment of the marginal benefits of the investment exceeds their marginal costs. Failing an exit examination could increase the "costs" of obtaining a high school diploma or attending college in several ways. First, it presents a structural barrier to graduation for students who cannot pass the examination on retest.[1] Second, preparing for and retaking the test may simply be a burden that some students are not interested in bearing.[2] Finally, failing the examination may affect students' perceptions of themselves and their academic abilities, which may in turn affect their ideas about the returns to additional schooling. It might also produce an emotional response as students who are told they are "failing" in mathematics and/or English Language Arts (ELA) become discouraged.

Over the past decade, researchers have begun to examine the consequences of failing high school exit examinations, using regression-discontinuity strategies to estimate the causal effect of barely failing an exit examination on subsequent student outcomes, including their future academic achievement, high school graduation, and early-career labor-market earnings.[3] The results have been mixed. Using data from Texas, Martorell (2005) found no effects of barely failing the 10th-grade exit examination early in the high school career on student high school graduation rates. In addition, Reardon et al. (2010) found no statistically significant effects on high school graduation rates (or most other outcomes) in either mathematics or ELA in California. By contrast, in earlier work in Massachusetts, Papay, Murnane, and Willett (2010) found that barely failing a first attempt at the 10th-grade high-stakes mathematics examination reduced the probability of on-time high school graduation by approximately 8 percentage points for urban, low-income students. Ou (2010) found quite similar results in New Jersey. These differences could reflect variation in the tests themselves, in the choice of cut scores, or in student populations across these different states.

In all of these states, students must pass multiple exit examinations to graduate from high school. Thus, all researchers in this area have faced situations in which these multiple

[1]Many states, including Massachusetts, explicitly attempt to limit this barrier by allowing students multiple retest opportunities. For example, some Massachusetts students retake the test more than six times. Nonetheless, some students still cannot meet the passing standard.

[2]They may judge that the extra time spent is not worthwhile; this may be particularly true for students whose scores are very far away from the passing cutoff.

[3]Note that the question of how failing an exit examination affects students is substantively different from the question of how imposing an exit-examination policy affects students. We focus on the first.

tests define a range of different treatment conditions. To date, researchers of high school exit examinations have addressed this challenge in one of two primary ways. Martorell (2005) incorporated information on all of the tests into a single analysis, compositing the students' test scores into a single forcing variable that represented a student's minimum test score across subjects. Because students in Texas must pass three exit tests to graduate, this approach enabled him to focus on students near the margin of passing the test in which they had performed the worst. However, this approach ignores potentially important differences in the effects of failing examinations in different subjects. Failing a mathematics examination may affect students substantially differently than failing an ELA examination does because, for example, remediation may be much easier in one subject than the other. Furthermore, this approach does not produce a consistent causal estimate unless certain restrictive conditions are met, such as the metric of measurement being identical—or at least equatable—across tests (Reardon & Robinson, 2012; Wong, Steiner, & Cook, 2012).

A second, less parsimonious approach involves analyzing the consequences of each exit examination separately. We pursued this strategy in our earlier paper (Papay, Murnane, & Willett, 2010), and fitted separate regression-discontinuity models for the mathematics and ELA forcing variables. We present results from a similar approach next, with a different outcome. The advantage is that, by pooling all students and focusing on a single cutoff, we can increase our statistical power. These analyses are also interesting substantively, as they represent the average causal effects of failing one examination across all students who score near the cutoff on that test, regardless of their score on the other. It is clear, however, that such separate analyses assume implicitly that treatment effects are homogeneous among students with different scores on the other examination; in other words, it assumes that failing the mathematics examination matters just as much for students who score "Advanced" in ELA as for those who fail the ELA test. Clearly, in a regime where students must pass both tests, failing two tests may be substantially worse than failing one.

Reardon, Arshan, Atteberry, and Kurlaender (2010) took this type of analysis one step further, examining whether the effect of failing one test in California depended on whether students passed the other.[4] They found no striking patterns. However, this approach may still obscure some heterogeneity in effects across levels of the other examination. For example, failing the ELA test may matter less for students who score very highly on the mathematics examination than for students who pass the mathematics test but score near the cutoff. In other words, students who feel confident about their mathematics ability might not be affected by the additional burden of passing the ELA examination. Alternately, failing the ELA examination might represent a greater barrier for these relatively high-performing students in mathematics because it may affect their self-esteem more. In short, we might not expect the effect of just failing on subsequent outcomes to be homogenous across scores on the other examination.

Consequently, these types of questions require us to model the impact of discontinuities in both forcing variables—the mathematics and ELA test scores—simultaneously. Approaches to analyzing such data have been described by Papay, Willett, and Murnane (2011), Reardon and Robinson (2012), and Wong et al. (2012). We distinguish between two cases. Wong et al. (2012) described a situation in which individuals' values on multiple forcing variables are used to assign them to a single treatment or control group. One example of this scenario would be if students who fail a test in either mathematics or ELA were assigned to a common summer school program. By contrast, Author (2011) and Reardon

[4]Berk and de Leeuw (1999) proposed and adopted a similar approach in their analysis of California's corrections system.

and Robinson described a situation more akin to the one we face in this article, in which multiple forcing variables are used to assign individuals to a range of treatments. In our case, we expect that the effects of barely passing the mathematics examination may be different than the effects of barely passing the ELA test. Although Wong et al. have several important lessons for the analyses we present, we follow more closely the approaches laid out by Reardon and Robinson and by Papay, Willett, and Murnane (2011).

Next we describe these approaches, explain the specific causal estimands that each produces, and examine their advantages and disadvantages. These new approaches provide more nuanced information about the consequences of student examination performance. We estimate the effect of barely passing both of these examinations on subsequent college enrollment and explore how the effect of just failing one test depends on scores on the other examination. In other words, we ask whether barely failing either (or both) the mathematics or ELA examination on the first try affects the probability of subsequent college enrollment for students on the margin of passing. If so, we then ask whether the effect of barely failing the ELA examination differs by mathematics test performance, and whether the effect of failing the mathematics examination differs by ELA test performance. In all cases, we focus on the probability that students attend college, an outcome that researchers of exit examinations have not explored.

ESTIMATING THE EFFECT OF DISCONTINUITIES IN MATHEMATICS AND ELA ON SUBSEQUENT COLLEGE ENROLLMENT

Site

The Commonwealth of Massachusetts is the site for our case study. Massachusetts is an interesting state to examine because its educational system is one of the best performing in the country and it has placed a high priority on educational reform. Since the Massachusetts Education Reform Act of 1993, which introduced standards-based reforms and state-based testing, Massachusetts has invested substantially in K-12 public education. This investment included a large increase in state funding of K-12 public education, as well as the creation of clearly defined academic standards and curriculum frameworks and a system of examinations aligned with the academic standards.

This focus on standards-based reform and these investments in public education appear to have paid off. The state has been praised for having the most rigorous academic standards in the country and state assessments that align closely with these standards (Finn, Julian, & Petrilli, 2006; Quality Counts, 2006). Furthermore, Massachusetts students are consistently among the nation's top performers on the National Assessment of Educational Progress examinations, and the state's National Assessment of Educational Progress performance has improved rapidly since the introduction of state testing (National Center for Education Statistics, 2008). Massachusetts is also a state with a relatively well-educated work force and a relatively strong economy. The state ranks first in the percentage of adult residents with a 4-year college degree (38%) and third in per-capita income ($51,254; U.S. Census Bureau, 2009).

In Massachusetts, beginning with the graduating class of 2003, students have had to pass the Massachusetts Comprehensive Assessment System (MCAS) tests in both mathematics and ELA in order to graduate from high school. Students take these examinations for the first time in the spring of 10th grade, and they receive their scores several months later. The state determines the passing—or "cutoff"—score confidentially each year, and all students

with scores that fall at, or above, the cutoff are deemed to have passed the test. Students have multiple opportunities to retake examinations if they fail.

Dataset and Sample

The Massachusetts Department of Elementary and Secondary Education has compiled a comprehensive database that follows students longitudinally through high school and tracks students who leave the system. It has provided us with access to these data. We have added information on college enrollments from the National Student Clearinghouse (NSC). These NSC data include students from nearly all colleges and universities (public and private, 2 year, and 4 year) in the United States. Student records are merged by the NSC using names and dates of birth. The match rate in Massachusetts approaches 95% in recent years (Dynarski, Hemelt, & Hyman, 2012).

In our analytic sample, we pool the 202,860 students who first took the 10th-grade MCAS examinations as sophomores in 2004, 2005, or 2006, respectively. For these three cohorts, passing the MCAS examinations was a requirement for graduation.[5]

Measures

Our primary outcome variable, named *COLL*, is dichotomous and indicates whether a student enrolled in college within 1 year of his cohort's high school graduation (= 1) or not (= 0). Recognizing that failing an exit examination may also affect the timing of students' college enrollments, we conduct supplementary analyses in which we redefine the outcome to record whether students enroll in college within 2 years of their cohort's high school graduation. Analyses of these two outcomes produce quite similar results, so we focus attention on the first.[6]

Our dataset also contains a record of student test scores; we focus our attention on scores from the student's first attempt on each of the exit examinations. To implement our regression-discontinuity approach, we create continuous forcing variables by centering students' raw test scores on the value of the corresponding minimum passing score, C.[7] On each of these recentered continuous predictors ($MATH^{C}$ and ELA^{C}), a student with a score of zero had achieved the minimum passing score. We also create dichotomous versions of the same predictors (*PASS_M* and *PASS_E*) to indicate whether a student's score met the passing standard on the relevant examination (= 1) or not (= 0).

Our dataset also includes information on selected exogenous demographic covariates, such as student race and gender, as well as indicators for whether the student was classified as limited English proficient, special education, or low income, or attended a high school in

[5]Less than 1% of all students, including those with serious special educational needs, may satisfy high school graduation requirements without taking the MCAS exit examinations. We exclude these students from our analysis.

[6]Results that examine the effect of barely passing one or more exit examinations on college-going within 2 years of the cohort's graduation are available from the first author on request.

[7]The cut scores differ by subject and year. For example, students had to earn 21 points to pass the mathematics examination in 2004 but only 19 points in 2005 and 20 points in 2006. The cut scores in ELA were 39, 38, and 35 points, respectively, across the 3 years. For more information on MCAS scoring and scaling, see the MCAS Technical Reports (Massachusetts Department of Education, 2002, 2005).

one of the state's urban school districts. We include these covariates, along with the fixed effect of cohort, in all of our statistical models in order to improve the precision of our estimation.

Analytic Approach: Causal Inference With Multiple Forcing Variables

As with any regression-discontinuity strategy, we seek to estimate the conditional mean of the outcome for individuals who fall into different "treatment" groups (e.g., those who just pass, or just fail, the examination) *at the cut score*. In the case of a single exit examination (e.g., in mathematics), then, our parameters of interest are therefore a pair of population means:

$$\begin{aligned} &\mu_r(COLL) = \lim_{x \to 0+} E[COLL \mid MATH_i^C = x] \\ &\text{and} \\ &\mu_l(COLL) = \lim_{x \to 0^-} E[COLL \mid MATH_i^C = x] \end{aligned} \tag{1}$$

Then the difference between these population means—which we represent by parameter τ—represents the causal effect of the treatment ("just passing the examination"), for students at the cutoff, where

$$\tau = \mu_r(COLL) - \mu_l(COLL) \tag{2}$$

In other words, because outcome *COLL* is dichotomous, parameter τ represents the difference in the population probability of attending college for otherwise equivalent students who pass (μ_r) and fail (μ_l) the test, *at* the margin of passing. Assuming that the cutoff on the forcing variable has been assigned exogenously, we can use a standard regression-discontinuity analysis to estimate this parameter (Murnane & Willett, 2011; Imbens & Lemieux, 2008; Lee & Lemieux, 2010).

However, students must pass exit examinations in both mathematics and ELA in order to graduate from high school. Because the state imposes its passing criteria rigidly, the discontinuities are sharp at both the mathematics and ELA score cutoffs. Thus, the four treatment conditions define four distinct regions in the two-dimensional space spanned by the forcing variables, ($MATH^C$, ELA^C). As illustrated in Figure 1, these regions are as follows:

1. If students pass Mathematics and pass ELA, they fall in Region A;
2. If students fail Mathematics and pass ELA, they fall in Region B;
3. If students pass Mathematics and fail ELA, they fall in Region C; or
4. If students fail Mathematics and fail ELA, they fall in Region D.

In our case, each of these four regions represents a substantively distinct treatment condition. Note that this situation is different from one in which multiple forcing variables assign individuals to a single treatment or control condition. For example, if failing either examination placed students in a uniform summer school program, other analytical strategies may be more appropriate (see Wong et al., 2012, for a more detailed discussion of this scenario).

Similar to the case with a single forcing variable, our parameters of interest then become the population conditional mean probabilities of attending college for individuals

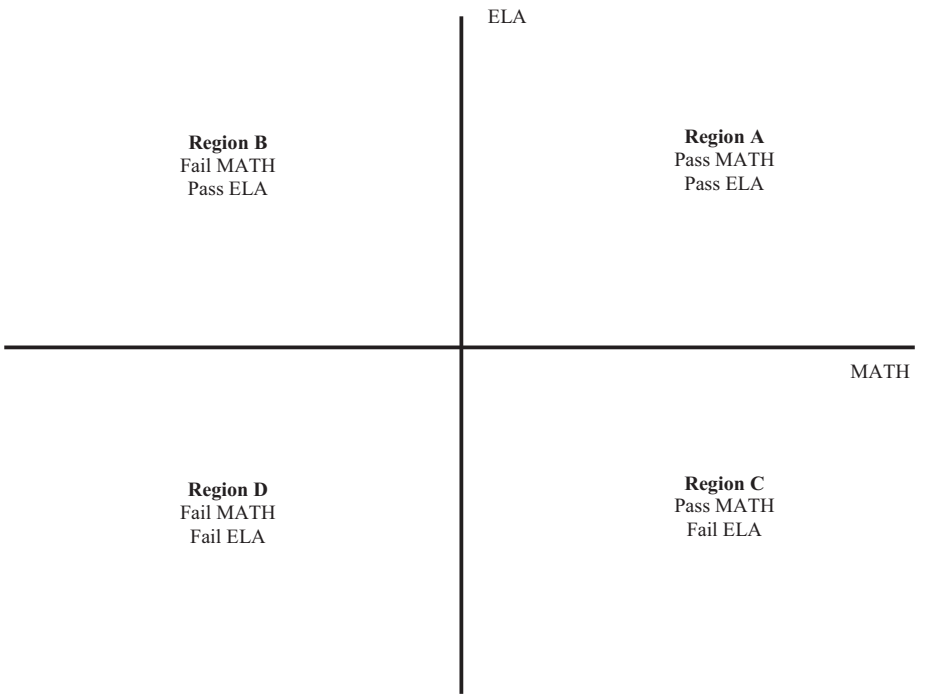

Figure 1. Treatment regions defined by two forcing variables, examinations in mathematics and English Language Arts (ELA).

in each treatment condition, at the relevant cutoff. For example, the causal effect of passing the mathematics examination instead of failing it for individuals scoring at the cutoff on the ELA test would be the difference between

$$\begin{aligned} &\mu_r(COLL) = \lim_{x \to 0^+} E[COLL_i \mid MATH_i^C = x, ELA_i^C = 0] \\ &\text{and} \\ &\mu_l(COLL) = \lim_{x \to 0^-} E[COLL_i \mid MATH_i^C = x, ELA_i^C = 0] \end{aligned} \tag{3}$$

We are interested in different comparisons among these treatments (or combinations of these treatments). For example, we can imagine nine different causal effects of interest (labeled CE), each of which refers to a different frontier (or section of a frontier) among the regions:

(CE-1) Average effect of barely passing the ELA examination (BA vs. DC)
(CE-2) Average effect of barely passing the mathematics examination (AC vs. BD)
(CE-3) Average effect of barely passing ELA, for students who fail math (B vs. D)
(CE-4) Average effect of barely passing ELA, for students who pass math (A vs. C)
(CE-5) Average effect of barely passing math, for students who fail ELA (C vs. D)
(CE-6) Average effect of barely passing math, for students who pass ELA (A vs. B)

(CE-7) Effect of barely passing both tests, compared to failing both, for students at the joint cutoff (A vs. D)[8]
(CE-8) Effect of barely passing ELA for students at any math score point (BA vs. DC at any point along the frontier)
(CE-9) Effect of barely passing math for students at any ELA score point (AC vs. BD at any point along the frontier)

It is important to note that we have ordered these causal effects from the aggregate to the specific. In other words, estimating effects CE-7, CE-8, and CE-9 would provide a much more nuanced picture of the impact of exit-examination performance on subsequent student outcomes than would CE-1 and CE-2. However, there are clearly trade-offs as estimating the more nuanced effects requires us to specify more complex regression-discontinuity models and, in general, to lose statistical power and precision as we spread the existing sample size out over statistical models of increasing complexity.

In all cases, we estimate the effect of "barely passing" the test, instead of "barely failing" the test for some group of students at the cutoff. In other words, we specify our models to estimate the coefficient on *PASS_M* and/or *PASS_E*. Thus, a positive coefficient indicates that barely passing the examination increases the probability of attending college. Importantly, we can interpret this same coefficient as suggesting that barely failing the examination reduces the probability of attending college. Thus, although we specify our models to estimate the effect of passing, we can interpret the parameter estimates as evidence about the effect of failing the test on college-going outcomes. As the two interpretations are mirror images, we use these terms interchangeably throughout our discussion.

Analytic Approach

Different regression-discontinuity (RD) approaches proposed in the literature enable us to estimate these different effects. Reardon and Robinson (2012) and Wong et al. (2012) provide a clear overview of these approaches, and we have proposed an alternative approach (Papay, Willett, & Murnane, 2011). Next we use Reardon and Robinson's terminology to describe several approaches, and we describe how we can fit these models and estimate their parameters.

Single Rating Score RD. Obtaining estimates of the average effect of barely passing one examination for all students near the cutoff (CE-1 and CE-2) is possible using standard, single forcing-variable RD approaches. Essentially, we ignore the role of the other forcing variable and examine the average predicted difference in subsequent outcomes for students who barely pass and barely fail one examination at the cut score. For example, to examine the effect of barely failing the mathematics examination on the probability that students attend college, we specify the following linear probability model[9] for the i^{th} student, in the

[8]Note that, for completeness, we could also be interested in estimating the effect of passing math and failing ELA, compared to passing ELA and failing math, for students near the joint cutoff (C vs. B). This is not a substantively interesting comparison in our example.

[9]Note that we have adopted to use a *linear probability*, rather than a *logistic* or *probit*, specification of the hypothesized relationship between our outcome—a dichotomous indicator of college attendance—and predictors. As noted by Angrist and Pischke (2008), in large samples, the linear probability specification provides unbiased (consistent) estimates of the underlying trends, while simplifying interpretation enormously. In addition, we use *local* linear-regression analysis—based only

vicinity of the cut score:

$$p[COLL_i = 1] = \beta_0 + \beta_1 PASS\_M_i + \beta_2 MATH_i^c + \beta_3(MATH_i^c \times PASS\_M_i) + \gamma' Z_i + \varepsilon_i \quad (4)$$

Regression parameters have their usual interpretation in terms of population differences in the probability of going to college per unit difference in the corresponding predictor, and ε is a residual with appropriate properties and distribution.

Here, cognizant of the potential sensitivity of any RD estimates to the functional form of the outcome versus forcing variable relationship, we follow procedures recommended by Imbens and Lemieux (2008), employing nonparametric smoothing using local-linear regression within a narrow bandwidth (h^*) to fit the hypothesized model and obtain estimates of its parameters. Imbens and Lemieux offered a method of cross-validation for determining the optimal bandwidth, which we follow. We have described our approach in greater detail, in an earlier publication (Papay, Murnane, & Willett, 2010). Effectively, in this example, the procedure provides a final estimate of the required causal effect by fitting a (locally) linear-regression model (such as Model 3) centered on the cut score, using only the subsample of students who fall within 2 points of the cutoff ($h^* = 2$).[10] As usual, we interpret parameter β_1 as the causal effect of passing the mathematics examination, instead of failing it, for students at the cutoff. We fit analogous models for scores on the ELA examination. Finally, following Lee and Card's (2008) admonition to account for the discrete nature of the scoring on the forcing variables, we estimate standard errors corrected for the clustering of participants at each discrete score point (or each combination of discrete mathematics and ELA scores in our multidimensional models) in all analyses in the article.

Frontier RD. Of course, this single rating-score approach is limited because it does not enable us to examine whether the effect of barely passing one examination on subsequent educational outcomes differs by student performance on the other. Reardon and Robinson (2012) described a "frontier" RD approach that involves fitting standard, single-forcing-variable RD models in different subsamples of students based on their performance on the other examination. Reardon et al. (2010) adopted this approach to investigate the effects of barely passing the high school exit examination in California. Here, to replicate their approach, we fit two separate regression models identical to those in (3), one using the subsample of students who pass the ELA examination and the other using the subsample who fail the test. Then, we fit two analogous regression models to estimate the effect of barely passing the ELA test, splitting the sample by students' mathematics performance. Fitting these four models enables us to estimate CE-3 through CE-6.

on observations within a narrow bandwidth—and so the linear specification of our statistical models is even more credible (as all trends become increasingly linear, locally, as the ranges of predictors are limited). As a result, when we replicate our analyses using the alternative logistic specification, our results are unchanged and almost identical to those we cite in the article. For example, after replicating our analysis under a logistic specification, we find that barely passing the mathematics examination increases the probability of college enrollment by 2.8 percentage points for prototypical students at the cutoff; the analogous effect in ELA in 4.6 percentage points. These correspond to the estimates of 2.8 percentage points (mathematics) and 4.5 percentage points (ELA) presented inTable 1.

[10]We carry out the cross-validation procedure described by Imbens and Lemieux (2008) to determine an optimal bandwidth separately for each subject. In all cases, we find an optimal bandwidth of 2 score points.

As Reardon and Robinson (2012) noted, one advantage of this approach is that we can use accepted, standard RD techniques, such as the nonparametric smoothing approach with local-linear regression that we have previously described. One disadvantage is that, by subsetting the data, statistical power is reduced. Furthermore, this approach does not enable us to address some substantive questions of interest, such as the effect of barely passing both examinations, instead of failing them.

Response-Surface RD. An alternative approach is to combine all these statistical models and estimate what Reardon and Robinson (2012) called the full multidimensional response surface, as follows:

$$p[COLL_i = 1] = \beta_0 + \beta_1 PASS\_M_i + \beta_2 PASS\_E_i + \beta_3 (PASS\_M_i \times PASS\_E_i) + \alpha' f(MATH_i^c, ELA_i^c) + \gamma' Z_i + \varepsilon_i \quad (5)$$

With the usual notion and where *f()* represents any continuous function of the forcing variables. In our example, we model this surface using a fifth-order polynomial in $MATH^C$, ELA^C, and the two-way interaction among their respective terms.[11]

The response-surface model enables us to replicate the estimation of the set of causal effects from the frontier RD approach, but simultaneously within a single statistical model. For example, in the response-surface model, parameter β_1 represents the population average effect of barely passing the mathematics examination, instead of failing it, for students who failed the ELA test (CE-5). Similarly, the parameter sum $\beta_1 + \beta_3$ represents the population average effect of barely passing the mathematics examination, instead of failing it, for students who passed the ELA test (CE-6).

One advantage of this response-surface approach is that it uses all available data, and thus the resulting parameter estimates are more precise. A serious disadvantage, though, is that, as Reardon and Robinson (2012) noted, we must rely on strong assumptions about the parametric functional form of the response surface to provide unbiased estimates of the treatment effect at each frontier. This has the effect of incorporating observations far from the cutoffs to project the relationship back to the cutoffs.

Both the frontier and the response-surface approaches enable us to explore treatment heterogeneity and to provide more nuanced interpretations of the causal effect of barely passing a high school exit examination than the single rating-score approach just described. However, they estimate explicitly the average effect of barely passing along each of the four frontiers. In other words, they estimate only CE-3 to CE-6. We might also be interested in an analysis that provides estimates of the effect of barely passing one examination by a more finer grained measure of a students' performance on the other examination. In other words, we might be interested in knowing whether the effect of barely passing the mathematics examination is greater for students scoring near the ELA cutoff than students scoring far from the cutoff. As Reardon and Robinson (2012) noted, we can modify both the frontier and response-surface approaches to estimate this heterogeneity directly by including statistical interactions between our treatment indicators and the forcing variables. Furthermore, in this context, we have proposed an approach (Papay, Willett, & Murnane, 2011) that uses nonparametric techniques to estimate these causal effects of interest more flexibly. We describe these three approaches next.

[11] In other words, we include $(MATH^C)^2$, $(MATH^C)^3$, $(MATH^C)^4$, $(MATH^C)^5$, $(ELA^C)^2$, $(ELA^C)^3$, $(ELA^C)^4$, $(ELA^C)^5$, $(MATH^C x ELA^C)^2$, $(MATH^C x ELA^C)^3$, $(MATH^C x ELA^C)^4$, and $(MATH^C x ELA^C)^5$. We test and reject that sixth-order polynomials contribute meaningfully to the prediction.

Frontier RD With Rating-Score Interactions. Here, we modify the frontier RD just described to include the statistical interaction between the relevant treatment indicator and the forcing variable of interest. Again, we subset our sample to include students in two of the four quadrants in Figure 1, and we fit a single-forcing-variable regression-discontinuity model using the techniques just described. For example, to estimate the effect of barely passing the mathematics examination for students at different ELA score points (CE-9), we fit models of the following form in two subsamples (students who passed the ELA examination and students who failed it):

$$\begin{aligned} p[COLL_i = 1] = \beta_0 + \beta_1 PASS\_M_i + \beta_2 MATH_i^c \\ + \beta_3 (MATH_i^c \times PASS\_M_i) + \beta_4 ELA_i^c \\ + \beta_5 (ELA_i^c \times PASS\_M_i) + \gamma' Z_i + \varepsilon_i \end{aligned} \tag{6}$$

Here, parameter β_1 represents the effect of barely passing the mathematics examination, instead of failing it, for students scoring *at the ELA cut score*, whereas parameter β_5 indicates whether this effect is larger or smaller for students further from the ELA cut score. For example, if we fit Model 5 in the subsample of students who passed the ELA examination, we would interpret a negative and statistically significant estimate of β_5 as evidence that students further from the cut score on ELA were less affected by barely failing the mathematics examination. It is important to note that we can trace out the effect of barely failing the mathematics examination at *each* ELA score point. For example, for students who score 5 points above the cutoff, we could test whether the linear combination $\beta_1 + 5^* \beta_5$ was zero, in the population.

One disadvantage of this frontier RD approach, though, is that it does not enable us to estimate one other key effect of interest: the effect of barely passing both examinations, instead of failing both. Doing so requires us to model the complete response surface in each quadrant, while allowing the effect of failing one examination to differ by the student's performance on the other. We outline two possible approaches: a modification of the response-surface RD approach that produces parametric estimates of these effects, and a nonparametric multidimensional RD approach that we have proposed (Papay, Willett, & Murnane, 2011) that uses local linear-regression instead of making complex functional form assumptions across the full range of data. Both approaches produce estimates of the effect of failing each examination at every level on the other test. As such, they enable us to estimate all nine of the causal effects previously listed, either directly or by using the methods of integral calculus to obtain average effect estimates over the range of interest.

Response-Surface RD With Rating-Score Interactions. First, we can modify the response-surface RD in Model 4 to include forcing-variable interaction terms. Again, we use a parametric fifth-order polynomial to model the response surfaces. However, we also interact our treatment indicators with the forcing variables, as follows:

$$\begin{aligned} p[COLL_i = 1] = \beta_0 + \beta_1 PASS\_M_i + \beta_2 PASS\_E_i + \beta_3 (PASS\_M_i \times PASS\_E_i) \\ + \beta_4 (MATH_i^c \times PASS\_ELA_i) + \beta_5 (ELA_i^c \times PASS\_MATH_i) \\ + \beta_6 (MATH_i^c \times PASS\_MATH_i \times PASS\_ELA_i) \\ + \beta_7 (ELA_i^c \times PASS\_MATH_i \times PASS\_ELA_i) \\ + \alpha' f(MATH_i^c, ELA_i^c) + \gamma' Z_i + \varepsilon_i \end{aligned} \tag{7}$$

For parsimony, we include only the linear versions of the forcing variables in the interactions listed, as we did in the frontier RD. In either case, we could also have included higher order terms to model a more complex relationship between the effect of passing one test and the test scores on the other.

Multidimensional RD. In Papay, Willett, and Murnane (2011), we describe in general terms a multidimensional regression-discontinuity approach that incorporates discontinuities in multiple forcing variables into a single analysis to estimate the causal effects of interest nonparametrically. This approach is quite similar conceptually to the response-surface model, in that we are estimating the relationship between our outcome and both of the forcing variables in each of the four treatment quadrants. However, we allow the relationship to differ in each of the quadrants and we estimate the relationship flexibly using nonparametric smoothing with local-linear regression. Thus, this approach does not require us to make strong functional-form assumptions over the whole range of the data, as in the response-surface model. The trade-off is that the approach requires a great deal of data and is more burdensome to implement. Here, we focus our attention on describing how we implement this approach using our current dataset.

We generalize Imbens and Lemieux's (2008) "nonparametric" approach from the single-variable case. We first choose an "optimal" joint bandwidth on the forcing variables (h^*_{MATH}, h^*_{ELA}) and then estimate the causal effect by conducting a local-linear regression analysis using this optimal bandwidth.

The primary challenge in implementing our approach comes in choosing the appropriate bandwidths (h^*_{MATH}, h^*_{ELA}) for our analysis. For each observation, at each point on the $(MATH^{\mathrm{C}}, ELA^{\mathrm{C}})$ grid, we fit a linear regression model—within an arbitrary bandwidth $(h_{\mathrm{MATH}}, h_{\mathrm{ELA}})$—to estimate a fitted value of the outcome at that point, as follows:

$$\hat{\mu}(MATH_i^C, ELA_i^C, h_1, h_2) = \hat{\gamma}_0 + \hat{\gamma}_1 MATH_i^C + \hat{\gamma}_2 ELA_i^C + \hat{\gamma}_3 (MATH_i^C \times ELA_i^C) \quad (8)$$

In each case, we use observations that fall within the appropriate region. Also, given that the regression-discontinuity approach explicitly focuses on estimates at the boundary of support (i.e., the cutoffs), we mirror this by estimating $\hat{\mu}(MATH_i^C, ELA_i^C, h_{MATH}, h_{ELA})$ as if it were a boundary point.[12] For a given bandwidth $(h_{\mathrm{MATH}}, h_{\mathrm{ELA}})$, we thus estimate a fitted probability of graduation for each observation. We then compare these fitted values to the observed values, summarizing them across the entire sample, using a generalized version of the Imbens and Lemieux (2008) cross-validation criterion:

$$CV_{COLL}(h_{MATH}, h_{ELA}) = \frac{1}{N}\sum_{i=1}^{N} (COLL_i - \hat{\mu}(MATH_i^C, ELA_i^C, h_{MATH}, h_{ELA}))^2 \quad (9)$$

[12]For example, for a student who fails both tests and for whom $MATH^{\mathrm{C}} = -3$ and $ELA^{\mathrm{C}} = -5$, we use only observations for which $MATH^{\mathrm{C}} < -3$ and $ELA^{\mathrm{C}} < -5$. By contrast, for students who pass both tests and for whom $MATH^{\mathrm{C}} = 10$ and $ELA^{\mathrm{C}} = 15$, we use only observations for which $MATH^{\mathrm{C}} \geq 10$ and $ELA^{\mathrm{C}} \geq 15$.

Our optimal joint bandwidth, h^*_{MATH} and h^*_{ELA}, is the pair of bandwidths that minimizes the CV criterion. In our case, this criterion reaches its minimum at ($h^*_{math} = 6, h^*_{ela} = 3$).[13]

After selecting an optimal bandwidth, we estimate the causal effect of passing the examinations using local linear-regression analysis. Some questions of interest involve the estimated effects near the joint cutoff; as such, we can derive identical estimates of these effects by fitting the requisite regression models in each region simultaneously (see Imbens & Lemieux, 2008). We can interpret this single model *parametrically* for observations local to the cut score and use the standard errors to conduct appropriate statistical tests. Thus, we specify a single statistical model with 16 parameters—an intercept, and slope parameters to accompany all 15 possible interactions among $MATH^C$, ELA^C, $PASS\_MATH$, and $PASS\_ELA$—and fit it using observations whose mathematics and ELA scores fall within one optimal bandwidth on either side of the relevant cut scores, as follows:

$$\begin{aligned} p[GRAD_i = 1] = {} & \beta_0 + \beta_1 PASS\_MATH_i + \beta_2 PASS\_ELA_i \\ & + \beta_3(PASS\_MATH_i \times PASS\_ELA_i) \\ & + \beta_4 MATH_i^c + \beta_5 ELA_i^c + \beta_6(MATH_i^c \times ELA_i^c) \\ & + \beta_7(MATH_i^c \times PASS\_MATH_i) + \beta_8(ELA_i^c \times PASS\_ELA_i) \\ & + \beta_9(MATH_i^c \times PASS\_ELA_i) + \beta_{10}(ELA_i^c \times PASS\_MATH_i) \\ & + \beta_{11}(MATH_i^c \times ELA_i^c \times PASS\_MATH_i) \\ & + \beta_{12}(MATH_i^c \times ELA_i^c \times PASS\_ELA_i) \\ & + \beta_{13}(MATH_i^c \times PASS\_MATH_i \times PASS\_ELA_i) \\ & + \beta_{14}(ELA_i^c \times PASS\_MATH_i \times PASS\_ELA_i) \\ & + \beta_{15}(MATH_i^c \times ELA_i^c \times PASS\_MATH_i \times PASS\_ELA_i) \\ & + \gamma' Z_i + \varepsilon_i \end{aligned} \tag{10}$$

As we discuss next, this model requires a large density of data near the joint cutoff to produce precise estimates.

It is important to note that, although this model provides estimates near the joint cut score, we need not restrict our attention to this area. Our approach also enables us to draw inferences about the causal effect of passing the mathematics examination across all levels of students' ELA scores (or vice versa). We again estimate these effects using local-linear regression, "sliding" the two-dimensional bandwidth smoothly along one or other of the forcing variables.

[13] Because data are less dense in the tails of the distribution, including observations that fall far from the cut score in the bandwidth estimation may lead us to select a larger bandwidth than is necessary. As a result, Imbens and Lemieux (2008) recommended deleting observations that fall beyond a certain quantile (δ) on either side of, and most remote from, the discontinuity before implementing the cross-validation procedure just described. Here, we set δ to 25% because the minimum passing scores fall within the tails of the joint distribution of the values of the forcing variables. We conduct this trimming process separately at each value of the forcing variables. In other words, at each value of $MATH^C$, we determine the 25th percentile of ELA scores for observations below the ELA cut score and the 75th percentile for observations above the ELA cut score. We follow a similar process, estimating relevant quantiles of mathematics score at each value of ELA^C. Then we exclude simultaneously all observations with $MATH^C$ or ELA^C scores more extreme than either of these sample quantiles, on either side of the cut score.

Finally, although this model enables us to estimate directly several key causal effects of interest (CE-7, CE-8, and CE-9), we can use it to recover all of the causal effects just described by using the methods of integral calculus to cumulate and average the estimates of interest across relevant ranges of the data (see Wong et al., 2012, for more discussion). For example, to estimate the causal effect of barely passing the ELA examination on average (CE-1), we can simply integrate all of the estimates of passing the ELA examination at individual mathematics score points across all mathematics scores present in the data. This same logic applies to all of the more complex models that we present.

RESULTS

We find strong evidence that barely passing an exit examination increases the probability that students will attend college. In Table 1, we present the estimated causal effects from our single-forcing-variable RD models (Equation 3). We estimate that barely passing the ELA examination increases the probability that students attend college by 4.5 percentage points ($p < .001$); in mathematics, the corresponding effect is 2.8 percentage points ($p < .001$). These effects are quite large given that only 27% of the students who score right at the ELA cut score and only 37% of students of the students who score right at the mathematics cut score attend college on-time. In Figure 2, we illustrate these estimated effects graphically, showing the fitted probability of attending college from our RD models imposed over the sample probability at each score point. The disruption in the underlying outcome/forcing variable trend at the cutoff is our estimate of the causal effect of barely passing the examination on attending college.[14]

These results suggest that exit-examination requirements do indeed have unintended consequences for students scoring near the margin. However, they obscure important interactions between the effect of barely passing one exit examination and student performance on the other test. As a result, we turn to results from the more flexible regression-discontinuity approaches. First, we examine whether the effect of passing one examination depends on students' performance level on the other test by implementing the frontier RD (Model 4) and response surface RD (Model 5) approaches. We present the estimated causal effects in Table 2.

[14] Again, we obtain similar results if we change our outcome to student college enrollments within 2 years of their cohort's high school graduation. Thus, it does not appear that the effect of exit examination performance on college-going is simply a matter of delaying college entry.

Table 1. Estimated causal effects of barely passing an exit examination on college enrollment for students at the margin of passing, separately in mathematics and English Language Arts (ELA) from the single rating-score regression-discontinuity model in (3)

Outcome	(CE-1) ELA	(CE-2) Mathematics
College Enrollment	0.045***	0.028***
	(0.002)	(0.000)
	$h^* = 2$	$h^* = 2$
	n = 8,673	n = 14,766

$^{\sim}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

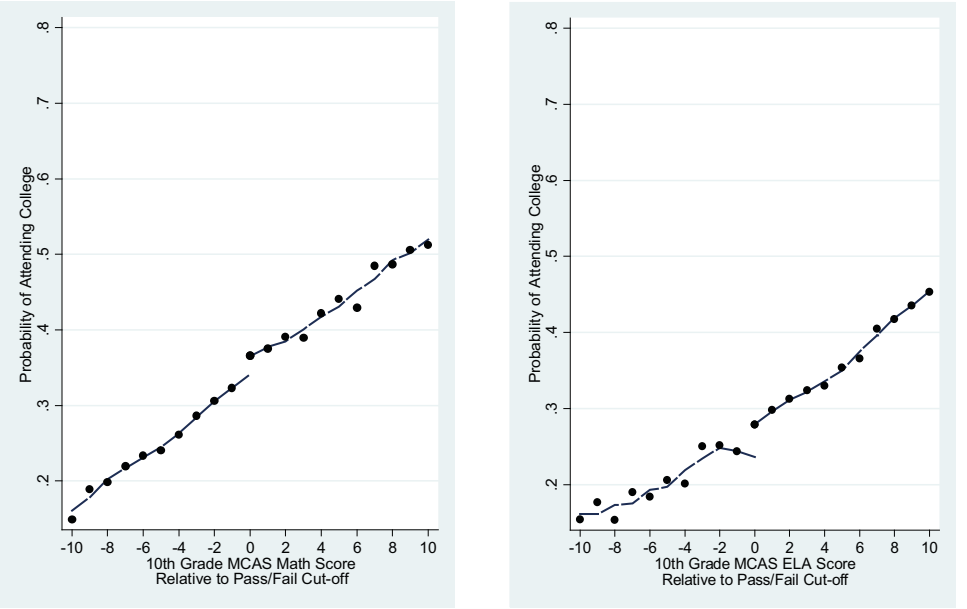

Figure 2. Smoothed nonparametric relationship (bandwidth = 2) between the fitted probability of attending college and scores on the mathematics (left panel) and English Language Arts (ELA; right panel) high school exit examinations, with the sample mean probabilities of attending college overlaid. *Note*. MCAS = Massachusetts Comprehensive Assessment System (color figure available online).

Table 2. Estimated causal effects of barely passing an exit examination on college enrollment for students on the margin of passing, separately in mathematics and English Language Arts (ELA), by their performance category on the other test, from the frontier RD model in (3) and the response-surface RD model in (4)

Causal Estimand	Frontier RD	Response Surface RD
(CE-3) Average effect of barely passing ELA, for students who fail math	0.023** (0.007) $h^* = 5$ n = 9,074	0.018* (0.009) n = 25,922
(CE-4) Average effect of barely passing ELA, for students who pass math	0.071*** (0.007) $h^* = 2$ n = 4,463	0.009 (0.011) n = 182,542
(CE-5) Average effect of barely passing math, for students who fail ELA	0.054*** (0.011) $h^* = 3$ n = 3,628	0.022* (0.011) n = 15,884
(CE-6) Average effect of barely passing math, for students who pass ELA	0.027*** (0.001) $h^* = 2$ n = 12,810	0.013~ (0.007) n = 192,580

Note. $^{\sim}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. RD = regression-discontinuity.

Interestingly, these two models produce somewhat different answers. The frontier RD estimate suggests that barely passing the ELA examination matters more for students who also pass the mathematics test ($\widehat{\beta} = 0.071$, $p < .001$), whereas barely passing the mathematics examination matters more for students who fail in ELA ($\widehat{\beta} = 0.054$, $p < .001$). These patterns suggest that the mathematics examination is more important as a "first hurdle" to college-going—for students who fail the ELA test, barely failing the mathematics examination substantially reduces their probability of enrolling in college. However, among students who pass the ELA test, whether they pass or fail the mathematics examination matters less ($\widehat{\beta} = 0.023$, $p = .006$). On the other hand, the ELA examination appears to be more important as a "second hurdle." For students who fail in mathematics, barely passing the ELA test increases their probability of college-going only modestly ($\widehat{\beta} = 0.023$, $p = .006$). However, for students who pass the math test, barely passing the ELA examination is quite important, increasing by 7.1 percentage points the fitted probability that students attend college.

By contrast, the estimated effects from the response-surface RD are much smaller. The mathematics results point in the same direction as in the frontier RD approach, but the estimated effects of failing the ELA examination do not. These inconsistencies suggest the importance of specifying these models correctly. One potential challenge with the response-surface RD approach is that we rely on observations far from the cutoff to contribute to the estimation of these effects, at the cut score. Thus, getting the functional form of the outcome/forcing-variable relationship correct is a real challenge, one that we may not have accomplished even with such a high-order polynomial. We assess this threat next, and find reason to believe that our results from the frontier RD are more robust.

For many applications, the types of questions addressed above will provide researchers with sufficiently nuanced information to examine the phenomenon of interest. However, we may also want to know whether these effects differ at specific points on the second forcing variable. To examine these questions, we modify the frontier RD and response-surface RD approaches to include interactions between forcing variables and indicators of passing. We also fit our multidimensional regression-discontinuity model. In the first row of Table 3, we present the estimated effects of barely passing both tests, compared to failing them, for students at the joint cutoff. These results clarify the important trade-offs made between these models. Most obviously, the frontier RD approach with rating-score interactions does not estimate this quantity. Of interest, we find discrepant results from the response-surface RD and multidimensional RD approaches. The response-surface RD provides a much more precise estimate ($\widehat{\beta} = 0.021$, $SE = 0.012$, $p = .079$), although again it relies on strong functional-form assumptions and observations far from the cutoff, which suggest that it is more susceptible to bias. By contrast, we find a much larger, but much less precise and not statistically significant, parameter estimate with the multidimensional RD approach ($\widehat{\beta} = 0.094$, $SE = 0.066$, $p = .159$). This approach uses nonparametric local-linear regression analysis and focuses on observations within a narrow window near the cutoff, reducing concerns of bias. But the limited sample size also reduces precision of the estimate substantially.

Although the joint cutoff is a point of particular interest, we are also interested in whether the effect of passing one examination differs across the entire distribution of scores on the other examination. In our case, we do not have sufficient data to make precise statements about these relationships; as such, we present the upcoming results as examples of the types of analyses that the different approaches can support. Because the frontier and response-surface RD approaches rely on modifications that treat these relationships parametrically, we can summarize the results in a single parameter estimate, as we do in

Table 3. Estimated causal effects of barely passing both exit examination on college enrollment for students on the margin of passing (row 1) and estimated slope of the relationship between the effects of passing one examination and student test scores on the other, from the frontier RD model in (5), the response-surface RD model in (6), and the multidimensional RD model in (9)

Causal Estimand	Frontier RD	Response Surface RD	Multi-Dimensional RD
(CE-7) Average effect of barely passing both, instead of failing both	N/A	0.021~ (0.012) n = 208,464	0.094 (0.066) n = 6,977
(CE-8A) Slope of relationship between effect of barely passing ELA and math score, for students who pass math	−0.001 (0.003) h* = 2 n = 4,463	−0.001 (0.001) n = 208,464	See Figure 3
(CE-8B) Slope of relationship between effect of barely passing ELA and math score, for students who fail math	0.000 (0.002) h* = 5 n = 9,074	−0.001 (0.002) n = 208,464	See Figure 3
(CE-9A) Slope of relationship between effect of barely passing math and ELA score, for students who pass ELA	0.002~ (0.001) h* = 2 n = 12,180	0.002~ (0.001) n = 208,464	See Figure 3
(CE-9B) Slope of relationship between effect of barely passing math and ELA score, for students who fail ELA	0.002 (0.002) h* = 3 n = 3,628	0.001 (0.002) n = 208,464	See Figure 3

Note. ~$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$. RD = regression-discontinuity.

the last four rows of Table 3. Here, we show how much the estimated effect of passing one examination differs with each test score point on the other examination; in other words, each parameter estimate represents the *slope* of the estimated effect of passing one test per unit difference in scores on the other test. For example, if we interpret the parameter estimate strictly, the second row suggests that the effect of barely passing the ELA examination is 1 percentage point less for students scoring 10 points above the mathematics cut score than students who just pass the mathematics test. However, the results in Table 3 show that there does not appear to be much of a relationship on either of the tests.

Because the multidimensional RD design uses a nonparametric approach, we cannot summarize the relationship in a single point estimate. As a result, we present the results from this analysis in Figure 3, where we plot the estimated causal effect of passing the mathematics examination, by ELA score.[15] In essence, we can see each point on the graph as a separate causal estimate of the effect of barely passing the mathematics examination. We also present the 95% confidence interval around each of these estimates. For example, we can interpret the parameter estimate where ELA = 10 on the *x*-axis as evidence that, for students scoring 10 points above the ELA cutoff, barely passing the mathematics

[15]We could construct an analogous plot showing the effect of barely passing the ELA examination by students' mathematics test score.

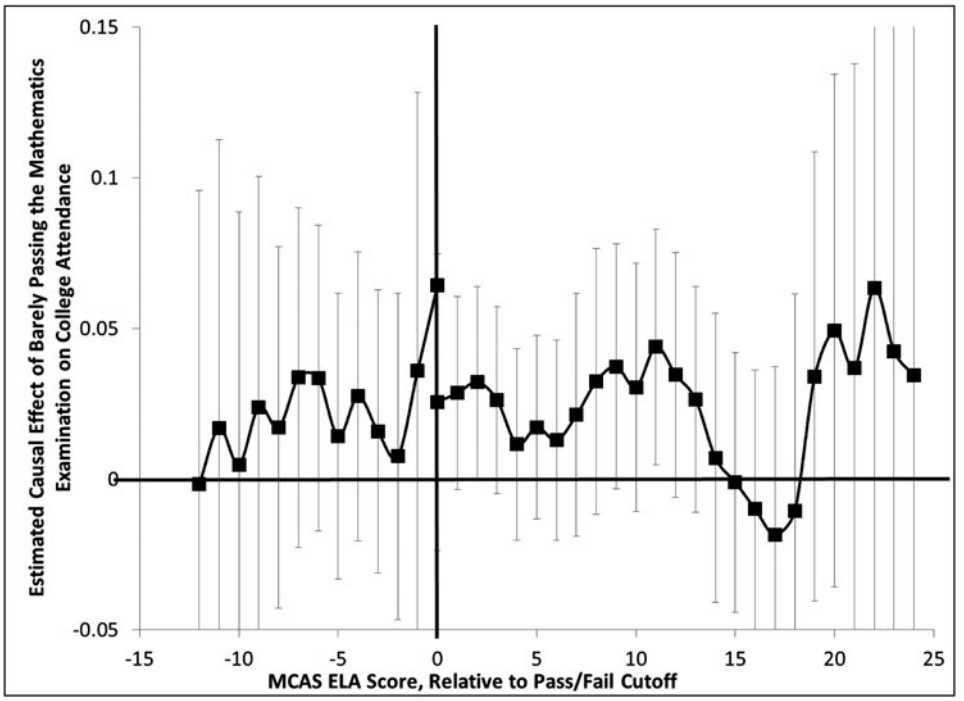

Figure 3. Graphical representation of the estimated causal effect of just passing the mathematics exit examination on college enrollment for students at the mathematics cut score, by English Language Arts (ELA) score, with corresponding 95% confidence intervals. *Note*. MCAS = Massachusetts Comprehensive Assessment System.

examination increases college-going by 3.1 percentage points. Again, these results are quite imprecise, and we cannot distinguish estimates from zero or from each other. In general, though, they echo the substantive findings of the frontier and response-surface RD approaches, particularly for students who pass the ELA examination (right of the cut score). Here, there is no clear relationship between ELA test performance and the effect of failing the mathematics examination on students' probability of attending college.

THREATS TO VALIDITY

Our ability to draw causal inferences from our regression-discontinuity design requires two important conditions to be met. First, treatment assignment—here whether students are classified as just failing, or just passing, one or more of the examinations—must be exogenous and applied rigidly to all students. All student characteristics, both observed and unobserved, must be smooth functions of the forcing variables around the cut scores. In other words, students must not be able to manipulate their positions relative to the cut scores on the forcing variable. In Massachusetts, this assumption holds because all students who score below the cutoffs fail the tests and all students who score above the cutoffs pass them. Furthermore, the minimum passing scores differ from year to year based on a complicated scaling formula and are determined *after* students take the tests; thus, students cannot manipulate their positions endogenously with respect to the cutoffs.

In addition to this prima facie evidence, however, we can assess whether the cut scores were imposed exogenously in several ways. First, we examine the distribution of students falling on either side of the cut scores and find no discontinuity in this density. Second, we examine whether there are apparent discontinuities at the joint cut scores in the average values of each of ten student-level covariates. Because we are seeking evidence of smoothness in these pretreatment covariates, we adopt the more flexible multidimensional RD approach. Treating each covariate now as an outcome, we use the same regression-discontinuity approach to estimate five relevant "effects" at the cut score.[16] Among these 50 estimates, we find only two for which we can reject the null hypothesis of no population effect at $p < .05$ and five at $p < .10$, just what we would expect from chance, with a 5% or 10% Type I error. Thus, we find no reason to doubt that the state has imposed the cut scores exogenously and consistently.

The second key assumption underpinning our regression-discontinuity analyses is that we must be able to estimate the relationship between the outcome and forcing variable adequately, at least in the immediate vicinity of the cut score. In the single forcing-variable case, we can examine the relationship visually, as in Figure 2. There, we can see clear visual evidence of a disruption in the smooth relationship at the cutoff. We also implement standard analyses to determine whether we have specified the relationship on either side of the cutoff using the appropriate functional form and whether our results are robust to bandwidth choice. We find that the general patterns described here hold (see the appendix for more detail).

Reardon and Robinson (2012) noted, though, that this is more difficult to assess with approaches that model the discontinuities in multiple forcing variables simultaneously. One key advantage of the single-variable approaches is that we can easily look for visual evidence of disruptions graphically. In a multidimensional RD model (or a response surface RD), this is much more difficult. One simple way to explore the robustness of these models is to examine whether they enable us to replicate the results from the single rating-score approach that uses well-defined RD methods. For example, we can compare the basic frontier and response-surface RD approaches by recognizing that the total effect of barely passing the ELA test is simply a weighted average of the estimated effects on students who pass the mathematics examination and those who fail it. From the frontier RD, we obtain an implied effect of barely passing the ELA test of 3.8 percentage points, which matches closely the figure of 4.5 percentage points in Table 1. However, the response-surface RD produces an implied effect of 1.0 percentage points. Thus, our results suggest that the frontier RD is likely superior to the response-surface RD to address these questions about simple heterogeneity.

We can apply similar logic to assessing the validity of the multidimensional RD approach, examining if this approach produces similar estimates to the simpler models. As previously described, we can estimate all of the causal effects of interest from this flexible specification by using integral calculus to cumulate and average the estimated effects across the range of test-score values of interest. For example, the estimated effect of barely passing the mathematics examination for students who fail the ELA test (CE-5) is simply the weighted average of estimated effects of passing the mathematics examination from the multidimensional RD approach at each of the ELA score points below the cutoff. We present these implied effects in Table 4, along with the estimated effect from our preferred simpler approach for each estimate. In nearly all cases, the multidimensional approach

[16]The effect of passing both tests (1), passing ELA for students who pass mathematics (2) and who fail mathematics (3), and passing mathematics for students who fail ELA (4) and pass ELA (5).

Table 4. Comparison of a range of estimated causal effects on college enrollment produced by our preferred model for each effect, along with the implied effect derived from the multidimensional RD model

Causal Effect	Preferred Model	Preferred Model Estimate	Multi-dimensional RD estimate
(CE-1) Average effect of barely passing ELA	Single rating-score RD	0.045	0.045
(CE-2) Average effect of barely passing math	Single rating-score RD	0.028	0.022
(CE-3) Average effect of barely passing ELA, for students who fail math	Frontier RD	0.023	0.024
(CE-4) Average effect of barely passing ELA, for students who pass math	Frontier RD	0.071	0.064
(CE-5) Average effect of barely passing math, for students who fail ELA	Frontier RD	0.054	0.019
(CE-6) Average effect of barely passing math, for students who pass ELA	Frontier RD	0.027	0.023
(CE-9) Average effect of barely passing both	Multi-dimensional RD	—	0.094

Note. RD = regression-discontinuity.

matches closely the estimated causal effect from the simpler model. This suggests that our model estimates the relationship between the probability of college attendance and the forcing variables near the cutoffs accurately. Nonetheless, we should view the results from the multidimensional model with somewhat more skepticism than results from single-variable approaches where we have greater statistical power and clearer tools—including visual analysis—to assess possible threats to validity.

DISCUSSION

Methodological Lessons

Recent increases in both the availability of large educational datasets and statistical computing power have opened up many possibilities for researchers to examine social-science questions using ever-more sophisticated analytic techniques. Situations where multiple forcing variables assign individuals to a range of different treatments are one such example. Here, we examine the effect of barely passing two exit examinations on the probability that students enroll in college. We explain that these two forcing variables define four treatments and nine possible treatment-control contrasts of interest. With additional forcing variables, the complexity becomes even greater. Thus, analysts seeking to investigate such situations need practical guidance to help select an approach. In this article, we have applied six different regression-discontinuity approaches using administrative data from the state of Massachusetts. Our results suggest that each approach has its strengths and weaknesses. Here, we highlight these differences and draw three main lessons for researchers.

First, in selecting an analytic approach, one must obviously be concerned with both bias and precision. As Reardon and Robinson (2012) noted, the response-surface RD approach uses the full range of data, making the estimates much more precise, but it relies on observations far from the cutoff to project estimates of the causal effects for students at the cutoff. In general, such approaches are not as robust as RD models that use nonparametric smoothing and estimate causal effects using only observations that fall within a narrow bandwidth around the cut score (see Imbens & Lemieux, 2008; Lee & Lemieux, 2010, for more discussion). Our analyses support this argument. This same logic applies to approaches that seek to estimate effects along several forcing variables. Thus, we prefer the single rating-score, frontier, and multidimensional RD approaches, even though they all capitalize on fewer observations and are less precise than the response-surface RD approach.

Second, in choosing between these three approaches, we recommend that researchers use the simplest technique that will enable them to answer their substantive question of interest. For example, if researchers want to know the average effect of barely passing the mathematics examination on college-going, there is no reason to adopt a frontier or multidimensional RD approach when the more straightforward single rating-score approach will suffice. On the other hand, certain questions require certain methods. For example, if researchers are interested in understanding how barely passing both examinations, instead of failing both, affects college attendance, the analytical options are limited to the response surface or multidimensional approaches. Thus, the substantive question should drive the approach used.

At the same time, substantive interest must be balanced with practical concerns about data availability. Each approach to examine how the effect of passing one examination differs by students' scores on another test involves fitting a complex statistical model with multiway interactions between test scores and indicators of passing and failing. As such, these models provide estimates that are less precise and require a great density of data, particularly near the cut scores, to provide substantively meaningful results. Even with more than 200,000 observations available in our Massachusetts data sets, our analysis produces estimates with relatively large standard errors because few observations are local to both cutoffs and they are not distributed uniformly across each of the four treatment conditions. For example, for most students in Massachusetts, passing the ELA examination is substantially easier than passing the mathematics test (only 2.0% of test takers pass the mathematics examination but fail the ELA test; by contrast, 6.6% pass the ELA test but fail the mathematics test). As seen in Figure 4, because the cut scores fall in the tails of the distribution, the density of data local to the joint cutoff is quite low. As a result, this analysis is underpowered, and we would need to be investigating large effects for all of our estimates to reach traditional levels of statistical significance.

It is important to keep in mind that the severity of this limitation depends in large part on the location of the cut score in the joint distribution of the forcing variables. For example, fitting our full model at a pseudo-discontinuity declared at the mean of each of the forcing variables, where the data distributions are much denser, produces much more precise estimates, with standard errors reduced by a factor of almost three. Thus, the particular data burdens of our approach arise in large part because the cut scores are in the tails of the forcing-variable distributions. These challenges may be less of a concern in other contexts. Thus, if the data requirements are met and the more nuanced analysis is substantively important, these techniques can provide valuable information for researchers and policymakers.

To summarize, we argue that a researcher's analytic approach should be driven by both substantive concerns and the properties of available data. To estimate average effects

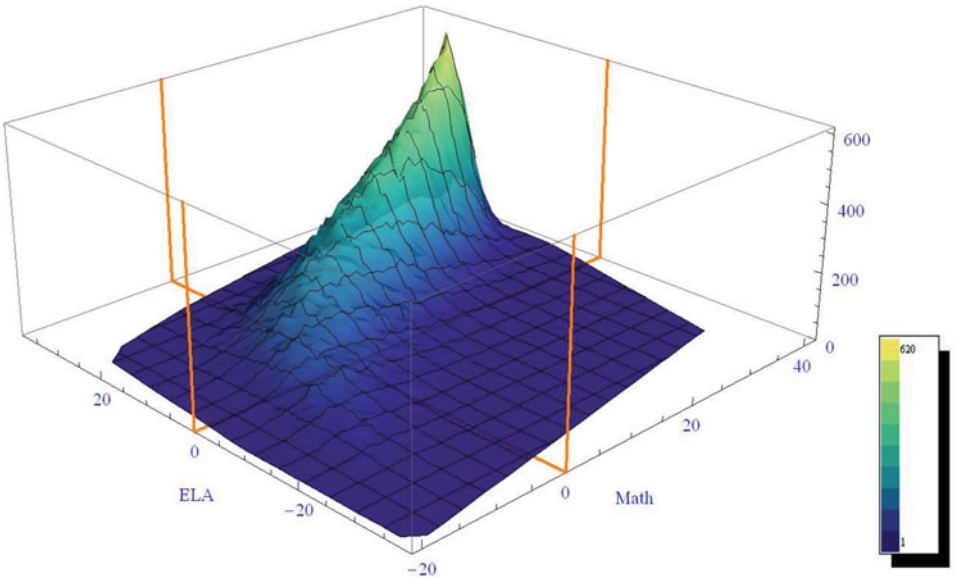

Figure 4. Surface plot of the empirical bivariate distribution of the sample data, by mathematics and English Language Arts (ELA) scores, centered on the corresponding pass–fail cut scores (color figure available online).

of passing a single test (CE-1 and CE-2), we recommend using nonparametric smoothing techniques to implement the single rating-score approach. To explore heterogeneity among students by their placement relative to the cutoff on the other test (CE-3 through CE-6), we recommend the frontier RD approach. Only in cases where researchers have a clear call to be interested in finer grained heterogeneity do we recommend the multidimensional RD model, which we argue is superior to the frontier RD or response-surface RD with rating-score interactions because these latter approaches must rely on stronger—and perhaps less credible—functional-form assumptions.

Finally, we should note that we have presented two extreme cases of examining treatment-effect heterogeneity. The basic frontier RD estimates the causal effect of barely passing one test, among all students who pass the other test or who fail it. The multidimensional RD approach estimates these same effects but at every score point on the other test. If there is reason to suspect that there are heterogeneous effects across levels of the other forcing variable but the multidimensional RD does not produce sufficiently precise answers, researchers could consider adopting a hybrid approach. In other words, rather than estimating the effect of barely passing the mathematics examination at every ELA score point, we could estimate the effect over small ranges of ELA scores. This intermediate approach may represent a useful balance of flexibility and statistical power in some applications.

Policy Implications

The factors that influence whether students continue to invest in schooling are not well understood. In this article, we examine one possible determinant: the increased costs of acquiring additional educational credentials that arise when students fail a high school exit

examination. Students who fail must spend time and effort studying for and retaking the examination, and failing may impose psychological costs on students because they may feel less confident about their academic abilities. We find that these costs are important drivers of student college-going decisions, at least for relatively low-performing students who score near the cutoff on the mathematics or ELA examinations.

We want to make clear two points relevant to the interpretation of our findings. First, we can say nothing about the overall effect of introducing an exit-examination policy. Simply having the requirement could improve student outcomes by making all students work harder throughout their school careers. On the other hand, students who believe they will never pass may drop out before even taking the first test, thereby reducing educational attainments. We focus on the consequences of passing or failing the exit examination in a state that has already implemented the policy. Second, we do not know whether these results represent the positive effect of barely passing the examination, the negative effect of barely failing it, or (more likely) some combination of the two.

Regardless of the mechanisms at play, two substantive lessons stand out. First, students who fail do not become more motivated to succeed in school; we find no evidence that failing the examination improves outcomes. Second, among relatively low performing but equally proficient students, barely passing or failing exit examinations influences the probability of college enrollment. These effects extend beyond high school. Students who fail their first test are not simply finding alternative routes into postsecondary education. Whatever the underlying mechanism, failing (or passing) a high school exit examination has long-lasting effects on students. Policymakers should attend carefully to these longer term effects.

These patterns raise several questions that provide important areas for future inquiry. First, the statistical power limitations of some of our analyses prevented us from examining heterogeneity across student types; given past work, we might expect these effects to concentrate more heavily in certain groups of students, such as low-income students in urban schools. Furthermore, we need to understand much more about the mechanisms that underlie these effects: Does barely passing the exit examination boost a student's self-concept (or does failing an exit examination reduce it)? Do students who barely fail perceive the burden of retaking the test to be too great, or do they simply not believe that they can achieve a passing score? Why do the effects of failing differ by subject, and by students' performance on the other examination? Answering these questions may increase understanding of teenagers' responses to the exit exam requirement that have become a part of the high school experience in the United States.

ACKNOWLEDGMENTS

The authors thank Carrie Conaway, the Associate Commissioner for Planning, Research, and Delivery Systems at the Massachusetts Department of Elementary and Secondary Education, for providing the data and for answering many questions about data collection procedures.

FUNDING

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E100013 to Harvard University and by the Spencer Foundation. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education or the Spencer Foundation.

REFERENCES

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Berk, R. A., & de Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association*, *94*, 1045–1052.

Center on Education Policy. (2010). *State high school tests: Exit exams and other assessments*. Retrieved from http://cep-dc.org/index.cfm?DocumentSubTopicID=8

Dynarski, S. M., Hemelt, S. W., & Hyman, J. M. (2012). *Data watch: Using National Student Clearinghouse data to track post-secondary outcomes* (Working paper).

Finn, C. E., Julian, L., & Petrilli, M. J. (2006). *The state of state standards*. Washington, DC: The Fordham Foundation. Retrieved from http://www.edexcellence.net/foundation/publication/publication.cfm?id=358

Goldin, C., & Katz, L. F. (2008). *The race between education and technology*. Cambridge, MA: Harvard University Press.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635.

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*, 655–674.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281–355.

Martorell, F. (2005). *Does failing a high school graduation exam matter?* Unpublished working paper.

Massachusetts Department of Education. (2002). *2001 MCAS technical report*. Retrieved from http://www.doe.mass.edu/mcas/2002/news/01techrpt.pdf

Massachusetts Department of Education. (2005). *2004 MCAS technical report*. Retrieved from http://www.doe.mass.edu/mcas/2005/news/04techrpt.pdf

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York: Oxford University Press.

National Center for Education Statistics. (2008). *State comparisons: National Assessment of Educational Progress (NAEP)*. Washington, DC: U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/nde/statecomp/

Ou, D. (2010). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review*, *29*, 171–186.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis*, *32*(1), 5–23.

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, *161*, 203–207.

Quality Counts. (2006). Quality Counts at 10: A decade of standards-based education. *Education Week*, *25*(17), 74.

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis*, *32*, 498–520.

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, *5*, 83–104.

U.S. Census Bureau. (2009). *The 2010 statistical abstract: The national data book*. Washington, DC: Author. Retrieved from http://www.census.gov/compendia/statab/rankings.html

Wong, V. C., Steiner, P. M., & Cook, T. D., (2012). Analyzing regression discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*. Advance online publication. doi:10.3102/1076998611432172

APPENDIX

One key challenge in all regression-discontinuity models involves specifying accurately the relationship between the outcome and the forcing variable. As previously discussed, our preferred models use local-linear regression analysis in a narrow bandwidth around the cutscore. Even though we have chosen an "optimal" bandwidth using a data-driven cross-validation procedure, sensitivity of our results to the bandwidth choice would cast doubt on the validity of our inferences. As all continuous functions are locally linear, functional form is less of a concern with this approach, provided that the bandwidth is sufficiently small. However, with larger bandwidths we might worry that the underlying linear trend does not fully capture the relationship of interest.

We examine the robustness of our estimates to bandwidth choice and functional form. To do so, we estimate our preferred model for causal effects (1) through (6) previously described using a range of bandwidths and higher order polynomials in the local-linear regression analysis. We present the results from these analyses in Table A1. For each causal effect of interest, we use our preferred approach. We fit 14 models, using bandwidths from 2 to 6 and linear, quadratic, and cubic polynomials.[17] Thus, each cell in the table represents a different estimate of the causal effect of interest.

Table A1. Estimated causal effects from our preferred approaches, across a range of bandwidths and functional forms

		Bandwidth				
Causal Effect	Specification	2	3	4	5	6
(CE-1) Average effect of barely passing ELA	Linear	0.045***	0.037***	0.018	0.018	0.015
	Quadratic	0.019***	0.056***	0.073***	0.039*	0.033 ∼
	Cubic	—	0.023***	0.032***	0.116**	0.073**
(CE-2) Average effect of barely passing math	Linear	0.028***	0.029***	0.022***	0.018**	0.025***
	Quadratic	0.033***	0.027***	0.033***	0.03***	0.018**
	Cubic	—	0.034***	0.023***	0.035***	0.043***
(CE-3) Average effect of barely passing ELA, for students who fail math	Linear	0.023*	0.017**	0.013*	0.023**	0.016*
	Quadratic	0.037***	0.03**	0.028***	0.003	0.023 ∼
	Cubic	—	0.042***	0.031**	0.065***	−0.006
(CE-4) Average effect of barely passing ELA, for students who pass math	Linear	0.071	0.062**	0.024	0.011	0.014
	Quadratic	0.005	0.087***	0.133***	0.092**	0.046
	Cubic	—	0.009*	0.021**	0.177**	0.185***
(CE-5) Average effect of barely passing math, for students who fail ELA	Linear	0.031**	0.054**	0.032**	0.022 ∼	0.017
	Quadratic	0.020**	−0.004	0.068*	0.054**	0.05**
	Cubic	—	0.004	−0.096***	0.052	0.049
(CE-6) Average effect of barely passing math, for students who pass ELA	Linear	0.027***	0.02**	0.017**	0.013*	0.025**
	Quadratic	0.037***	0.039***	0.025**	0.025***	0.004
	Cubic	—	0.041***	0.064***	0.034**	0.052***

[17]Note that we cannot include a cubic term in models where we limit ourselves to a bandwidth of 2 points.

In general, the results are quite stable across bandwidths and polynomials. Given that we use a data-driven local-linear regression analysis in most of our approaches and that the bandwidths are quite small, we are less concerned about the need to include higher order polynomial terms. Thus, we find the first row of each panel of the table to be most informative. Furthermore, given that our optimal bandwidths are all quite small, we find the estimates from bandwidths of 2, 3, and 4 to be most informative. Of course, across all 14 estimates we do see some variation. In general, though, the results appear to be fairly robust. It is important to note that not only do the patterns of estimates appear to be relatively consistent, with few exceptions, but the average of the estimates in each of these panels matches closely the effect estimates from our preferred models.