

NBER WORKING PAPER SERIES

THE CONSEQUENCES OF HIGH SCHOOL EXIT EXAMINATIONS FOR STRUGGLING LOW-INCOME URBAN STUDENTS  
EVIDENCE FROM MASSACHUSETTS

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

Working Paper 14186  
<http://www.nber.org/papers/w14186>

NATIONAL BUREAU OF ECONOMIC RESEARCH  
1050 Massachusetts Avenue  
Cambridge, MA 02138  
July 2008

The authors thank Carrie Conaway, the Director of Planning, Research, and Evaluation of the Massachusetts Department of Elementary and Secondary Education, for providing the data and for answering many questions about data collection procedures. Participants in the May 1, 2008 NBER economics of education workshop provided helpful comments. Financial support was provided by the U.S. Department of Education Institute for Education Sciences (Grant Number R305A080127) and the Harvard Graduate School of Education Dean's Summer Fellowship. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2008 by John P. Papay, Richard J. Murnane, and John B. Willett. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Consequences of High School Exit Examinations for Struggling Low-Income Urban Students:  
Evidence from Massachusetts

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

NBER Working Paper No. 14186

July 2008

JEL No. I21

**ABSTRACT**

The growing prominence of high-stakes exit examinations has made questions about their effects on student outcomes increasingly important. We take advantage of a natural experiment to evaluate the causal effects of failing a high-stakes test on high school completion for the cohort scheduled to graduate from Massachusetts high schools in 2006. With these exit examinations, states divide a continuous performance measure into dichotomous categories, so students with essentially identical performance may have different outcomes. We find that, for low-income urban students on the margin of passing, failing the 10th grade mathematics examination reduces the probability of on-time graduation by eight percentage points. The large majority (89%) of students who fail the 10th grade mathematics examination retake it. However, although we find that low-income urban students are just as likely to retake the test as apparently equally skilled suburban students, they are much less likely to pass this retest. Furthermore, failing the 8th grade mathematics examination reduces by three percentage points the probability that low-income urban students stay in school through 10th grade. We find no effects for suburban students or wealthier urban students.

John P. Papay  
Harvard Graduate School of Education  
Cambridge, MA 02138  
jppapay@fas.harvard.edu

Richard J. Murnane  
Graduate School of Education  
Harvard University  
6 Appian Way - Gutman 409  
Cambridge, MA 02138  
and NBER  
richard\_murnane@harvard.edu

John B. Willett  
Graduate School of Education  
Harvard University  
6 Appian Way  
Cambridge, MA 02138  
John\_Willett@Harvard.Edu

# **The Consequences of High School Exit Examinations for Struggling Urban Students: Evidence from Massachusetts**

## ***I. Introduction***

As part of standards-based educational reforms introduced over the past two decades, many states have implemented exit examinations that students must pass in order to earn high school diplomas. Advocates argue that such examinations create incentives for students to work at learning important cognitive skills. By certifying that high school graduates have mastered the state-defined academic content standards, the examinations may also increase the economic value of a high school diploma (Evers & Walberg, 2002). Opponents of these tests suggest that they put unnecessary stress on students and encourage them to drop out of high school. They also argue that such tests place the greatest burden on the very groups who are already struggling in the educational system, such as low-income and special needs students (Thomas, 2005; Jones, Jones, & Hargrove, 2003). Because high school graduation is associated with many positive life outcomes, the question of how high-stakes testing affects high school completion rates is important to educational policymakers.

While past research has focused primarily on the overall effects of imposing high-stakes accountability on aggregate student outcomes, we look at the consequences of exit examinations on individual students within a high-stakes testing regime. Capitalizing on a natural experiment, we examine the causal impact of failing the statewide 10<sup>th</sup> grade mathematics examination on the probability of on-time high school graduation. The natural experiment stems from the state's practice each year of determining a minimum passing score, thereby dividing a continuous performance measure into two categories – pass and fail. We use a regression discontinuity methodology to examine the consequences of being assigned to each of these categories for

students of similar proficiency near the cutoff. Our data come from Massachusetts, a state that has earned a national reputation for rigorous content standards and English Language Arts (ELA) and mathematics assessments that are well aligned to the standards, and whose students have made substantial progress under standards-based reform. Thus, we examine these effects for students under an existing high-stakes accountability system.

An exit examination can prevent students from graduating from high school in three ways: fear of failing may cause them to drop out before taking the test; failing the examination may cause them to drop out before re-taking it; and failure to pass even after multiple attempts may prevent graduation. We refer to these mechanisms as *Fear of Failure*, *Discouragement*, and *Repeated Failure*. We conduct a variety of analyses to explore the extent to which each of these mechanisms affects academically struggling students.

We find that, for equally able low-income, urban students near the cutoff, failing the 10<sup>th</sup> grade mathematics exit examination – as opposed to passing it – reduces the probability of on-time graduation by eight percentage points. In contrast, failing the test does not reduce the probability of on-time graduation for wealthier urban students or suburban students on the margin of passing. Thus, the combination of low family income and urban schooling makes students particularly susceptible to the effects of failing. Importantly, we cannot distinguish here between a negative effect of failing the examination and a positive effect of passing it – students who fail the test may be disappointed with their performance and drop out of school, while students who pass may be encouraged and persist in school. Regardless, the practice of dividing students with essentially the same ability into two categories by this dichotomous cut score has an impact on student outcomes; this effect poses an important challenge for urban districts.

Furthermore, for a typical low-income urban student on the margin of passing the 8<sup>th</sup>

grade mathematics examination, failing that test also reduces the probability of persisting through 10<sup>th</sup> grade by three percentage points, providing some evidence that *Fear of Failure* may play a role in students' decisions. Interestingly, we find that failing the 10<sup>th</sup> grade ELA examination does not affect the probability of graduation for low-income urban students on the margin of passing.

We supplement these causal conclusions with descriptive analyses that explore possible sources of these effects for urban students with low family income. Here, we focus on students who fail the mathematics exit examination when they first take it at the end of grade 10, exploring their persistence and success on retests. At each retest opportunity, more than 80% of students who fail continue to retake the test. Massachusetts students show remarkable persistence, but relatively few students exhaust all of their retest opportunities. Instead, of the students who fail and never pass the examination, nearly two-thirds stop taking retests and drop out of school, presumably because their poor test performance has discouraged them. Here, we find important differences for urban, low-income students compared to their suburban peers. Among students with the same initial test scores, low-income urban students who fail the statewide mathematics examination at the end of the 10<sup>th</sup> grade are just as likely as suburban students to retake the test, but they are much less likely to pass on retest. Differences in academic support could explain this pattern.

In Section II, we provide a brief discussion of standards-based reforms, their development in Massachusetts, and past research on the effects of high-stakes testing. In Section III, we explain our data sources, measures, and analytic strategy. Here, we justify our ability to make causal claims from these data. In Section IV, we detail our main findings. In Section V, we perform several sensitivity analyses to verify the robustness of our results. We conclude with

a discussion of our findings and the questions they raise for policy-makers.

## ***II. Background and Context***

### Standards-Based Educational Reforms and High-Stakes Testing

In the years since the 1983 publication of *A Nation at Risk*, the standards-based reform movement has gained momentum and exerted substantial influence on state and federal education policy. While the details of these reform efforts vary greatly from state to state, common components include specification of content standards in core academic subjects and regular testing to monitor student progress toward mastering these standards. In addition to developing accountability structures for schools, many states have begun attaching consequences for students to their performance on the state-wide examinations. Currently, 25 states have or are phasing in examinations, typically in English language arts (ELA) and mathematics, that high school students must pass in order to graduate (Center on Education Policy, 2007). In most states, including Massachusetts, students first take these exit examinations as 10<sup>th</sup> graders. Students who fail typically have multiple opportunities to retake the examination before graduation.

Critics of high-stakes examinations argue that they may lead some students to drop out of high school (Thomas, 2005; Jones, Jones, & Hargrove, 2003). A 1999 *National Research Council* report cites qualitative work suggesting that “graduation tests pose no threat to most students, but, among those who fail them, they increase a sense of discouragement and contribute to the likelihood of dropping out” (Heubert & Hauser, p. 175). Any policy that causes students to drop out of school has substantial consequences because high school graduation remains a gateway into better paying jobs and post-secondary education. Employers recognize and reward

the skills that college graduates possess, especially the ability to engage in non-routine problem-solving and to communicate effectively (Levy & Murnane, 2004). Because it causes students to complete less education, dropping out also reduces students' quality of life in a variety of dimensions, including reduced health, wealth, and happiness (Oreopoulos, 2007), increased criminality (Lochner & Moretti, 2004), and increased mortality rates (Lleras-Muney, 2005).

Given the importance of high school completion and the possible negative consequences of high-stakes testing, many scholars have explored whether exit examinations reduce graduation rates. This work has taken two main forms: some researchers have examined the effect of imposing high-stakes testing on aggregate student outcomes, while others have focused on the relationship between an individual student's performance on the test and that student's probability of graduating in states with high-stakes testing regimes.

Much early work examining aggregate outcomes used correlational evidence; Clarke, Haney, & Madaus (2000) review this literature and conclude that "high stakes testing programs are linked to decreased rates of high school completion." Exploiting variation in exit examination policies across states and/or over time, some recent work provides at least tentative support for these correlational conclusions (Reardon & Galindo, 2002; Warren, Jenkins, & Kulick, 2006; Nichols, Glass & Berliner, 2006). In contrast, Carnoy & Loeb (2002), Greene & Winters (2004), and Carnoy (2005) find no relationship between state accountability policies, including high school exit examinations, and high school completion rates. Some recent work suggests that exploring aggregate patterns may obscure heterogeneity in effects for different groups of students. Dee and Jacob (2006) find increased dropout rates only for urban and minority students, while Jacob (2001) finds similar patterns only for the lowest achieving students.

Research that examines the relationship between individual student performance on exit

examinations and high school completion remains much less common. Using data from the Florida Minimum Competency Test from 1987-91, Griffin & Heidorn (1996) find a relationship between student performance and drop-out rate only for students with high GPAs. While the authors control for selected student characteristics, the results cannot be interpreted causally because it is likely that students who fail the examination differ from those who pass in critical unobserved dimensions. Griffin & Heidorn also focus on the impact of a minimum competency test, which differs substantially from the current incarnation of state-mandated high school exit examinations. Cornell, Krosnik & Chang (2006) examine a group of students who were wrongly informed that they had failed the Minnesota high-stakes examination. Most of these students reported some negative academic impact of “failing” this test.

Martorell (2005) provides causal estimates of the effect of failing a high school exit examination on high school graduation, using a regression discontinuity analysis similar to the one we employ in this paper. He finds no effect of failing the Texas exit examination on high school graduation for students who barely failed. This finding holds for every examination until the very last administration of a student’s senior year. As students run out of testing opportunities, failing the examination does prevent them from graduating because they cannot satisfy state requirements.

We do not address the overall consequences of standards-based accountability in Massachusetts. Instead, like Martorell (2005), we look at how this policy plays out in a state committed to standards-based reform. In other words, we look at the effects of dividing a continuous measure of student proficiency into two categories – pass and fail – at an arbitrary cut score. We extend Martorell’s research in several respects. Most importantly, we look for (and find) heterogeneous causal effects. In Massachusetts, examining only aggregate impacts masks a

substantial effect for low-income urban students; as a result, we focus our analyses on this group. Second, we examine additional mechanisms by which exit examinations may decrease high school graduation, including the possibility that students drop out even before taking the 10<sup>th</sup> grade test. Third, we conduct descriptive analyses that shed light on the sources of the heterogeneous causal impacts. Finally, we make use of data from a state quite different from Texas.

### The Massachusetts Context

In the 15 years since the Massachusetts legislature passed the *Massachusetts Education Reform Act* of 1993, the state has invested more than one billion dollars per year in additional funding for K-12 public education. These investments have borne considerable fruit. For example, a 2006 study by the Fordham Foundation praised the Massachusetts academic standards as the most rigorous in the country (Finn, Julian, & Petrilli, 2006).<sup>1</sup> A 2006 report by *Education Week* concluded that the state-wide tests used to assess the English language arts and mathematical skills of Massachusetts students (part of the Massachusetts Comprehensive Assessment System (MCAS)) were well aligned with the state's demanding academic standards. While this report gave an average grade of B- to the standards and accountability systems developed by states, it gave the Massachusetts system an A (*Quality Counts*, 2006).

Most importantly, Massachusetts students are doing well and have improved markedly on the National Assessment of Educational Progress (NAEP) examinations in recent years. In 2007

---

<sup>1</sup> This same Fordham Foundation report, *The State of State Standards 2006*, pointed out that the Massachusetts standards were exceptional. In contrast, “two-thirds of schoolchildren in America attend class in states with mediocre (or worse) expectations for what their students should learn” (Finn, Julian, & Petrilli, 2006).

Massachusetts' 4th graders ranked first nationwide on the NAEP reading and mathematics tests and second nationwide on the writing test. The state's 8<sup>th</sup> graders ranked first in mathematics, tied for first in reading, and third in writing on the NAEP tests (NCES, 2008). Furthermore, since the introduction of state testing under standards-based reform, the state's NAEP performance has improved rapidly. As Figure 1 shows, Massachusetts 8<sup>th</sup> graders not only far exceed the national average, but their performance has increased much more rapidly than the national average. Thus, it is in the context of a system that has brought about significant accomplishments that we examine the consequences for students of failing the MCAS examination.

FIGURE 1 ABOUT HERE

Massachusetts began administering the MCAS mathematics and ELA examinations in 1998. For the class of 2003, the 10<sup>th</sup> grade tests became high-stakes exit examinations. Students must pass both tests in order to receive a high school diploma.<sup>2</sup> The state allows students to take the tests without time constraints and to retake them repeatedly if they fail, attempting explicitly to make the MCAS as minimal a barrier to graduation as possible.<sup>3</sup> Critics, however, claim that even with these safeguards, the examinations do indeed prevent students from graduating.

---

<sup>2</sup> As the state imposed high stakes on the MCAS for students, student performance has risen dramatically. The overall passing rate jumped from 49% to 68% in the year of the policy shift. Currently, 87% of students pass the test. The state estimates that this effect represents approximately a 0.5 standard deviation increase in student test performance simply as a result of imposing the requirement to pass (Conaway, personal communication, 2008).

<sup>3</sup> The state has a performance appeals process in place that allows students to demonstrate their proficiency in alternate ways. It also offers alternative assessments to certain students. Only 314

Of the nearly 70,000 students who took the 8<sup>th</sup> grade mathematics examination in 2002, 76% went on to graduate on time in Massachusetts in 2006. We can partition those students who did not graduate on time into two groups – those who did not persist to take the 10<sup>th</sup> grade examination (9%) and those who took the 10<sup>th</sup> grade test but did not graduate two years later (15%). Thus, most students who did not graduate left the system after taking the 10<sup>th</sup> grade examination. We focus first on this population and return to the group who dropped out before 10<sup>th</sup> grade later in the paper.

That students who passed the 10<sup>th</sup> grade MCAS examination on their first attempt graduate at greater rates than students who fail is not surprising – all students must pass the test to graduate. Of the 66,347 students in the 2006 graduating cohort who took the 10<sup>th</sup> grade MCAS mathematics examination for the first time in 2004, 87% passed on their first try. However, students who failed faced substantial risk of dropping out: only 50% of them went on to graduate on time, compared to 90% of the students who passed.

While striking, this descriptive pattern does not confirm that the exit examinations pose a barrier to graduation. A student's MCAS scores are associated with a variety of other characteristics, such as academic proficiency, motivation, and access to educational resources, that also affect their probability of graduation. As a result, we would expect students who fail the examination to drop out at greater rates, even in the absence of any testing requirement. The direct relationship between MCAS score and the graduation rate among students who did pass the 2004 test provides evidence for this conclusion. Among these students, 73% who just passed

---

of the state's 57,000 graduates in 2006 satisfied the requirement using either of these two alternative routes.

graduated on time, compared to 98% of students with a perfect score.<sup>4</sup> Thus, a challenge in our current study involves disentangling the effects of failing the examination from the effects of student ability and other background characteristics related to test performance.

Conceptually, we would like to take students who scored identically, right at the pass/fail cut score, and randomly assign them to either a “pass” or a “fail” condition. This assignment process would render them equivalent in expectation on all observable and unobservable characteristics prior to treatment, allowing us to identify any differences in the ultimate outcome (high school graduation) as a causal effect of simply failing the examination, rather than of earning lower scores. Such an experiment is, of course, both impossible and unethical. However, we can take advantage of the state’s exogenous imposition of a minimum passing score to provide a natural experiment from which we can draw equivalent causal conclusions. By examining students with nearly identical MCAS performance, but just on either side of this exogenously-assigned cutoff, we can interpret any differences in their graduation outcomes as the causal effect of failing the examination for these students “on the margins” of passing (Shadish, Cook, & Campbell, 2002).

### Research Questions

We first examine the effect of failing the 10<sup>th</sup> grade examination, paying particular attention to impacts on low-income, urban students. We then attempt to explore the *Fear of Failing* effect that may arise as students predict they will not pass the 10<sup>th</sup> grade test and drop out

---

<sup>4</sup> For scores above the passing standard, the estimated correlation between the raw MCAS mathematics score and the proportion of students who graduate on time is 0.965, suggesting a very strong positive linear relationship between MCAS performance and probability of on-time high school completion.

before even taking it. We cannot identify this effect cleanly, but we can get some sense of its magnitude by examining student performance on the 8<sup>th</sup> grade test. Finally, we look at students who fail their 10<sup>th</sup> grade test, examining their persistence and success on retests. We explore whether students who fail and drop out do so because of *Discouragement* (students give up and drop out after failing one or more of the examinations) or through *Repeated Failure* (after exhausting the available retest opportunities, students still have not satisfied the testing requirements). Specifically, we address three primary research questions:

- RQ1. Does failing the high school exit examination as a 10<sup>th</sup> grader make students on the margin of passing less likely to graduate from high school on time?
- RQ2. Among students who fail the 10<sup>th</sup> grade exit examination and do not graduate, is the primary mechanism one of *Discouragement* or *Repeated Failure*?
- RQ3. Does failing the 8<sup>th</sup> grade test cause students on the margin of passing to drop out before taking the 10<sup>th</sup> grade examination?

### ***III. Research Design***

#### **Data Sources**

The Massachusetts Department of Education has compiled a comprehensive database that tracks students longitudinally throughout high school, allowing for clear description of student graduation outcomes. For the 2006 graduating cohort, the records contain each student's MCAS mathematics and ELA test results, demographic characteristics, and status at cohort graduation, including whether the student graduated, dropped out, is still enrolled, transferred out, was expelled, or any of eleven possible outcomes. This dataset allows for much more precise estimation of the probability of high school completion than do previous studies, and it permits

investigation of the direct link between student performance on high-stakes tests and graduation outcomes at the individual level.

Our dataset includes 83,892 student records from across the state of Massachusetts. To analyze the effect of failing the 10<sup>th</sup> grade examination (our first research question), we focus on members of the 2006 graduating cohort who first took the 10<sup>th</sup> grade mathematics MCAS examination as sophomores in 2004 and for whom the examination was a high-stakes test. This sample includes students who entered the state between 8<sup>th</sup> and 10<sup>th</sup> grade and consequently missed the 8<sup>th</sup> grade examination. Our final sample for addressing the first research question includes 66,347 students.<sup>5</sup> For our third research question, we use the 69,127 students in the same cohort who took the 8<sup>th</sup> grade mathematics examination. This sample includes students who dropped out of school before 10<sup>th</sup> grade.

### Measures

To address our first research question, we created a dichotomous outcome variable, named *GRAD*, that indicates whether the student graduated from a Massachusetts high school in Massachusetts in 2006 (1=graduated on-time in Massachusetts; 0 otherwise). Districts report the values of individual student graduation outcomes to the Department of Education using the state's *Student Information Management System* (SIMS). Note that students can be coded as zero

---

<sup>5</sup> The state identifies slightly fewer than 3,000 students (less than 5% of the total sample) who are not in the “final 2006 cohort,” meaning that they moved out of the state before high school graduation. Using only the 63,361 individuals in the “final cohort” does not alter our results. We include the full sample to account for any effects the high-stakes examination has on student mobility.

either for dropping out of school, for moving out of state before graduation, or for continuing in high school without graduating. In Section V, we explore the sensitivity of our results to this outcome definition. We created several additional outcome variables for the descriptive analyses that we used to address our second research question. For students who failed the examination, we created dichotomous outcomes that indicate whether the student retook the test (*RETAKE*) and whether they passed this retest (*RETAKE\_PASS*). Finally, to address our third research question, we created another dichotomous outcome measure, named *TAKE10th*, that indicates whether a student who took the 8<sup>th</sup> grade mathematics examination persisted in school to take the 10<sup>th</sup> grade test (1=persisted to take the test; 0 otherwise).

The dataset contains a record of scores from every MCAS mathematics and ELA examination that each student took from 8<sup>th</sup> grade<sup>6</sup> through high school graduation. The state reports raw scores, scaled scores, and performance level for each test. A scaled score of 220 qualifies as passing, with a different performance rating each 20 points, as follows: (a) 200 to 218: Failing, (b) 220 to 238: Needs Improvement, (c) 240 to 258: Proficient, and (d) 260 to 280: Advanced. Since multiple raw scores translate to a single scaled score, we use raw scores in our analyses in order to preserve fine-grained performance differences on the test.<sup>7</sup> For the 10<sup>th</sup> grade mathematics examination, raw scores ranged from 0 to 60; students who earned more than 20

---

<sup>6</sup> Technically, students took the middle school ELA examination in 7<sup>th</sup> grade and the mathematics examination in 8<sup>th</sup> grade. For simplicity, we refer to these examinations as the “8<sup>th</sup> grade” tests.

<sup>7</sup>The state reports reliabilities of 0.92 for mathematics and 0.89 for ELA. For more information on MCAS scoring and scaling, see the MCAS Technical Reports (MA DOE, 2002, 2005).

points passed the test.<sup>8</sup> To implement our regression discontinuity approach, we centered students' raw scores by subtracting out the value of the corresponding minimum passing score. On these re-centered continuous predictors, *MATH* and *ELA*, a student with a score of zero had achieved the minimum passing score. We also created a dichotomous predictor, *PASS*, to indicate whether the student passed the examination (1=student passed; 0 otherwise).

The dataset also includes the values of several key control predictors, such as student race and gender as well as dichotomous variables indicating whether the student was classified as limited English proficient (*LEP*), special education (*SPED*), low-income (*LOWINC*), attending a high school in one of Massachusetts's 22 urban school districts (*URBAN*), or appearing in the 10<sup>th</sup> grade sample without an 8<sup>th</sup> grade test score (*NEWSTUDENT*).<sup>9</sup> Each of these indicators is coded 1 for those who belong to the category, and 0 otherwise. Overall, 26% of the students attended urban schools and 28% of students were identified as low income. Low-income students tended to cluster in urban schools: 63% of urban students lived in poverty, compared to just 16% of suburban students.

### Data Analyses

We address our first and third research questions by conducting identical regression discontinuity analyses with the relevant outcome variable. We describe below the analyses that we use to address our first research question, which concerns the impact of just failing the 10<sup>th</sup>

---

<sup>8</sup> For the 8<sup>th</sup> grade mathematics test, students had to score 22 points to pass, and for the 10<sup>th</sup> grade ELA examinations the minimum passing score was 39.

<sup>9</sup> Some of these students moved into the state after 8<sup>th</sup> grade, while others simply had missing 8<sup>th</sup> grade test scores. Because we cannot distinguish between these two groups, we cannot interpret this variable as a pure indicator of new students to the state.

grade mathematics examination on the probability of on-time high school graduation. To explore whether just failing the 8th grade mathematics test reduces persistence to 10th grade (our third research question), we replace outcome *GRAD* by outcome *TAKE10th*.

Under conditions that we discuss below, we can analyze data from our natural experiment – using the regression discontinuity strategy first proposed by Thistlethwaite & Campbell (1960) – to make such causal inferences *for students at the margins of passing*.<sup>10</sup> Because the probability that a student passes the examination goes unequivocally from zero to one at a single cut score, the discontinuity is sharp.

The internal validity of our regression discontinuity analyses – and consequently our ability to make unbiased causal inferences about the impact of exit examinations – relies on several critical assumptions about the relationship between student MCAS score and graduation. Later in the paper we describe our efforts to verify that these assumptions are fulfilled. If so, the magnitude of the discontinuity in the outcome provides an unbiased estimate of the causal impact of failing the examination for students at the cut score. Thus, we obtain an estimate of the average treatment effect for students on the margin of passing.

We estimate the effect of failing the examination as a difference in the probability of on-time graduation between students scoring at the cutoff who just passed ( $\gamma_{pass}$ ) and just failed ( $\gamma_{fail}$ ).<sup>11</sup> In our analyses, we use observations above the cut score to estimate  $\gamma_{pass}$  and observations below the cut score to estimate  $\gamma_{fail}$ . Because we do not know the precise functional form of the relationship between MCAS score and the probability of graduation, we

---

<sup>10</sup> For a more detailed description of the regression discontinuity approach see Shadish, Cook, & Campbell (2002).

<sup>11</sup> Technically,  $\gamma_{pass} = \lim_{MATH_i \rightarrow 0^+} [P(GRAD_i = 1) | MATH_i]$  and  $\gamma_{fail} = \lim_{MATH_i \rightarrow 0^-} [P(GRAD_i = 1) | MATH_i]$

model this continuous relationship using a nonparametric smoothing process to estimate  $\gamma_{pass}$  and  $\gamma_{fail}$ . A further complication arises as our parameters of interest –  $\gamma_{pass}$  and  $\gamma_{fail}$  – are estimated at boundary points. As standard nonparametric smoothing strategies have poor boundary properties, Hahn, Todd, & Van der Klaauw (2001) recommend estimating these limits with local linear regression.<sup>12</sup>

Our implementation of nonparametric smoothing using local linear regression follows closely the recommendations of Imbens and Lemieux (2007).<sup>13</sup> We conduct our nonparametric smoothing within a linear probability specification of the standard regression discontinuity design. Specifically, at each MCAS score point, we estimate a linear regression function using only observations within a narrow bandwidth,  $h$ , around the point to predict the probability of graduation for each observation. As we move this bandwidth through our data range, we therefore generate locally predicted values at each MCAS score point; linking these estimates together creates the requisite smoothed nonparametric regression line. Here, the extent of the smoothing depends on the choice of bandwidth,  $h$ . Because we can only make causal claims about the effect of failing for students at the cut score, in our later analyses we focus attention on the single locally-linear regression analysis that centers on the cut score and estimates  $\gamma_{pass}$  and  $\gamma_{fail}$ . In this regression, then, we use only observations within bandwidth  $h$  on either side of the cut score, as follows:<sup>14</sup>

---

<sup>12</sup> Fan (1992) shows that, unlike most nonparametric smoothing techniques, local linear regression does not require boundary modifications.

<sup>13</sup> Ludwig and Miller (2007) use a similar strategy. Our approach differs in our choice of a rectangular rather than a triangular kernel for the non-parametric smoothing; however, Imbens & Lemieux (2007) argue that “more sophisticated kernels rarely make much difference” (p. 16) and instead recommend assessing robustness to different bandwidth choices, as we do in Section V.

<sup>14</sup> We estimate robust (Huber-White) standard errors to account for both the clustering of students within schools and heteroscedasticity in the dichotomous outcome.

$$p(\text{GRAD}_i = 1) = \beta_0 + \beta_1 \text{MATH}_i + \beta_2 \text{PASS}_i + \beta_3 (\text{PASS}_i \times \text{MATH}_i) + \varepsilon_i \quad (1)$$

for the  $i^{\text{th}}$  individual. While our nonparametric smoothing approach does not, by definition, return parameter estimates,<sup>15</sup> we can interpret the estimates from this single locally-linear fit in (1); these estimates represent the instantaneous slopes and intercepts for students at the cut score. In this model, parameter  $\beta_2 = \gamma_{\text{pass}} - \gamma_{\text{fail}}$  represents the causal effect of passing the 10<sup>th</sup> grade MCAS mathematics examination on the population probability of on-time high school graduation for students at the cut score. If its estimated value is statistically significant and positive, then we know that classifying a student as *passing* the high-stakes test at the cut score, as opposed to *failing* it, causes the student's probability of graduating from high school to *increase* discontinuously.

Our nonparametric procedure requires that we choose a suitable bandwidth,  $h$ , for the smoothing procedure and consequently for defining the region around the discontinuity in which we fit and interpret the model in (1). In our analyses, we select an optimal bandwidth,  $h^*$ , using all of our data by applying the cross-validation procedure described by Imbens & Lemieux (2007). Essentially, this procedure determines the bandwidth that minimizes the mean squared error in the predicted boundary points, leading to an optimal tradeoff of bias and precision for the estimation of  $\gamma_{\text{pass}}$  and  $\gamma_{\text{fail}}$ .<sup>16</sup> In our analyses, we obtain an optimal bandwidth of between four

---

<sup>15</sup> For example, the overall relationship between MCAS score and probability of graduation cannot be represented by a single slope throughout the data range.

<sup>16</sup> In other words, we determine a predicted probability of graduation ( $\hat{G}RAD_i(h)$ ) for each observation  $i$  using only observations within  $h$  points to the left of  $MCAS_i$  for students who failed and to the right of  $MCAS_i$  for students who passed the examination. We determine the mean squared error of these predictions across the entire sample. We then systematically vary the bandwidth,  $h$ , choosing as  $h^*$  the value of  $h$  that minimizes this mean squared error. More formally,  $h^* = \arg \min_h \frac{1}{N} \sum_{i=1}^N (\hat{G}RAD_i(h) - GRAD_i)^2$ . Because our ultimate objects of interest are the parameter estimates at the cut score, Imbens & Lemieux recommend excluding observations

and six raw score points depending on the model specification, as indicated below. However, in sensitivity analyses described in Section V, we show that our main conclusions are robust to the choice of bandwidth.

We extend this simple model in several ways. First, we include a vector of selected student background covariates ( $\mathbf{X}_i$ ) to improve precision and to eliminate small sample biases that result from including observations not immediately at the cut score (Imbens & Lemieux, 2007). Second, because our primary outcome is a dichotomous predictor that indicates whether the student graduates from high school on time, we replicate our analysis by specifying the probability of on-time high-school graduation as a logistic function of predictors. Here, we limit our analysis to those observations that fall within a narrow window around the cut score. For consistency with our earlier nonparametric smoothing, we choose a window whose width extends the optimal bandwidth of  $h^*$  on either side of the cut score. Again, we systematically vary this window width in Section V in order to test the robustness of our findings.<sup>17</sup>

Finally, we also examine the impact of test failure on high school graduation for particular groups of students, including urban students from low-income families. We do this in two ways. First, we add all possible interactions between predictors  $PASS_i$ ,  $MATH_i$ ,  $LOWINC_i$ , and  $URBAN_i$ , up to and including the four-way interaction among the predictors, to our regression equation in (1). Second, we fit separate regressions for each subgroup. As we find nearly identical results, we present this more parsimonious approach. Again, our main results

---

in the tails from the cross-validation determination. As data are less dense in the tails, including these observations may lead to over-smoothing. As a result, we eliminate the 10% of the observations on either side of, and most remote from, the cutoff.

<sup>17</sup> In preliminary analyses, we investigated whether higher-order non-linear polynomial specifications of  $MATH$  score were required within the logistic model, including quadratic and cubic polynomial specifications. These specifications did not lead to improvements in model fit, within the narrow regression discontinuity window that we have selected for the analysis, and so we present results from the more parsimonious linear specification here.

here derive from a single local linear regression analysis that incorporates only observations within an optimal bandwidth,  $h^*$ , on either side of the cut-off.

For two reasons we focus on the statistically significant impact of just passing/failing the MCAS mathematics test on high school graduation for low-income urban youth. First, the educational challenges facing these students have received national attention. Consequently, understanding the impact of high-stakes testing on the academic prospects for struggling low-income urban students is especially relevant to educational policy formulation. Second, the data currently available to us are insufficient to support exploration of other interesting questions, such as the effect of just failing the 10<sup>th</sup> grade MCAS test on urban special education students. We plan to examine additional subgroup effects in future research after we have increased our sub-sample sizes by pooling data across multiple graduation cohorts.

To address our second research question, we conduct analyses in which we explore *why* failing the 10<sup>th</sup> grade MCAS mathematics test reduces the probability of high school graduation for low-income urban students, but not for their wealthier or suburban peers. However, we interpret these results only descriptively because the additional analyses cannot support unbiased causal inference. In these descriptive analyses, we explore patterns of test-taking persistence and success for students who fail, in order to see whether low-income urban students are less likely than wealthier or suburban students to retake the examination or to pass their first retest. Here, we fit probit models of the following form on the sample of students who failed the 10<sup>th</sup> grade mathematics examination:

$$\begin{aligned}
 RETAKE_i = & \beta_0 + \beta_1(URBAN_i \times LOWINC_i) + \beta_2URBAN_i + \beta_3MATH_i + \beta_4ELA_i \\
 & + \beta_5PASS\_ELA_i + \tilde{\alpha}' \mathbf{X}_i
 \end{aligned} \tag{2}$$

for the  $i^{th}$  student.<sup>18</sup> In our model, our principal research interest focuses on the parameter sum,  $\beta_1 + \beta_2$ , which represents the difference between low-income urban students and suburban students in the probability of retaking the test. By including mathematics and ELA test scores and whether the student passed the ELA examination in the model, we explicitly compare students with the same proficiency on both the mathematics and ELA examinations.

We conduct similar analyses to examine retesting success, replacing the outcome  $RETAKE_i$  with  $RETAKE\_PASS_i$ . One challenge with this approach involves the differential effects of measurement error on retest success for low-income, urban students and their suburban peers who fail.<sup>19</sup> For these analyses, we use a truncated sample that generates estimates of retest success conditional on having failed the examination. Because the passing score is further at the tail of the distribution for suburban students than for urban, low-income students,<sup>20</sup> suburban students who fail are more likely than low-income, urban students to have performed so poorly on the test merely by chance. As a group, then, suburban students who fail are likely to do better mechanically on the retest than low-income, urban students, regardless of any increase in true proficiency. In other words, truncating our sample to focus on students who failed their first examination induces a correlation between the error term and our indicators of group membership, preventing us from obtaining unbiased estimates of  $\beta_1$ . Simulation results, presented in Appendix A, confirm that suburban students will outperform low-income, urban

---

<sup>18</sup> In preliminary analyses, we found that low-income and wealthier suburban students were indistinguishable from one another in terms of their probability of retaking the examination or of passing their first retest. By omitting the main effect of dichotomous predictor  $LOWINC_i$  from the hypothesized model, we implicitly treat all suburban students, regardless of family income, as the reference group.

<sup>19</sup> The authors thank Steven Rivkin for pointing out this issue and for his helpful suggestions for addressing it.

<sup>20</sup> The passing score is at the 32<sup>nd</sup> percentile for urban, low-income students but just the 9<sup>th</sup> percentile for suburban students.

students on the retest mechanically, without any difference in true proficiency. As recommended by Hanushek & Rivkin (2006), we resolve this issue by exploiting the fact that the mathematics and ELA test administrations occur on different days and measurement error on the two tests is uncorrelated. Given the strong correlation between mathematics and ELA performance, we use 10<sup>th</sup> grade ELA scores to instrument for 10<sup>th</sup> grade mathematics scores and adopt a two-stage least squares estimation strategy. This approach breaks the link between measurement error and group membership. Simulation results suggest that this approach successfully resolves the issue (see Appendix A).

#### ***IV. Findings***

##### **(1) Effect of failing the high-stakes exit examination on high school graduation**

Passing the 10<sup>th</sup> grade MCAS mathematics examination increases the probability that a low-income, urban student on the margin of passing will graduate from high school on-time by eight percentage points ( $p=0.015$ ). Given that 26% of low-income, urban students who just pass the exam do not graduate on time, this effect is quite substantial. We find no such effects for wealthier urban students or for suburban students, regardless of family income. Thus, it is the interaction of low family income and an urban environment that appears to render students, on average, more susceptible to the effects of failing. In Table 1, we present parameter estimates and approximate  $p$ -values from our local linear regression analyses using observations that fall within our “optimal” window of  $h^*$  on either side of the cut score. Models 1a and 1b present our findings for all students from equation (1), with time-invariant student demographic controls, by subgroup.

TABLE 1 ABOUT HERE

To interpret the estimates presented in Table 1 more easily, we present the fitted nonparametrically smoothed relationship between graduation and MCAS mathematics score for low-income urban students from our preferred specification in Figure 2.<sup>21</sup> For these low-income urban students at the margin, passing the examination substantially increases their subsequent probability of graduation. Visually, this effect appears as an interruption in the underlying smooth relationship between the probability of graduation and the MCAS mathematics score at the cut score. For perspective, we have included the sample mean probabilities of on-time graduation at each MCAS score level.

#### FIGURE 2 ABOUT HERE

The effects for wealthier urban and suburban students are not statistically significant. However, the point estimates indicate that wealthier students on the margins of passing who just fail have a slightly greater probability of on-time graduation than students who just pass. This seemingly counterintuitive pattern could stem from efforts by schools with ample resources to focus attention on the relatively few students with failing MCAS scores. Recent research by Neal and Schanzenbach (2007) lends some support for this claim; the authors find that, in the Chicago Public Schools, teachers face and respond to incentives to focus instruction on students who seem likely to improve their performance on the high-stakes examination.

#### (2) Persistence and success in retesting among students who fail

Overall, the 8,269 students who failed the mathematics MCAS on their first try in 2004 showed remarkable persistence in retaking the examination. Nearly 89% took the examination at least one more time and, of these students, 68% went on to pass the test at some point in high

---

<sup>21</sup> We can also recover the fitted relationship between graduation and MCAS mathematics score for the three other categories of students (wealthier urban, low-income suburban, and wealthier suburban). However, as our analyses show no effects on these groups, we decide to focus on the relationship for low-income urban students.

school. On average, students who never passed the examination retook it twice before giving up. As the sample histogram in Figure 3 illustrates, on each retest, approximately 35% of the students passed. Among those who failed each retest, most students (85 to 90 percent) decided to retake it yet another time. Although not shown, the numbers of students pursuing retests declines precipitously after the fourth retest: only 113 students retook the examination a fifth time, and only 7 took a sixth retest. Thus, very few students took advantage of all retest opportunities.

#### FIGURE 3 ABOUT HERE

Among students who failed their first test, we find evidence for both *Discouragement* and *Repeated Failure*. Here, we examine students whose initial test scores placed them within one bandwidth below the cut score but who never pass a retest. More than two-thirds of these students stop taking retests at some point and do not attempt the March 2006 examination, the last retest before the cohort's graduation. Presumably, these students become discouraged and dropped out of school. However, one-third of these students persist to the March 2006 retest. Over 85% of these students have taken at least four retests, showing remarkable persistence. For these students, *Repeated Failure* appears to be the mechanism at play as they exhaust all of their retest opportunities but cannot satisfy the graduation requirement. These patterns support the Massachusetts Department of Education's claim that most students have ample opportunities to retake the examination.

Table 2 includes parameter estimates and approximate  $p$ -values from fitting the models specified in equation (2) to predict the probability that students who failed the 10<sup>th</sup> grade mathematics examination retake and pass the first retest. In Figure 4, we present the fitted probability of retaking the examination (top panel) and passing the first retest (bottom panel) as a

function of initial mathematics test score.<sup>22</sup> Figure 4 illustrates that, among students with the same predicted MCAS scores on the initial tests, low-income urban students are no less likely than suburban students to retake the mathematics examination. However, low-income urban students are nearly ten percentage points less likely to pass this retest than suburban students with the same initial scores ( $p < 0.001$ ).

TABLE 2 ABOUT HERE

FIGURE 4 ABOUT HERE

### (3) Effect of failing the 8<sup>th</sup> grade examination on persistence to 10<sup>th</sup> grade

Although the 8<sup>th</sup> grade examination does not carry high stakes for students, performance on the test is clearly related to the probability that students remain in school through 10<sup>th</sup> grade. We present results in Table 3 from a regression discontinuity analysis of this outcome. For low-income urban students on the margin of passing the 8<sup>th</sup> grade mathematics test, failing reduces the probability of continuing in school and taking the 10<sup>th</sup> grade MCAS examination by three percentage points ( $p = 0.16$ ). While this effect is not statistically significant in the model estimated with optimal bandwidth, we arrive at nearly identical, but more precise and statistically significant results using a slightly larger bandwidth. Because only eleven percent of low-income urban students who just pass the examination leave the system before 10<sup>th</sup> grade, this three percentage point decline is noteworthy. In Figure 5, we illustrate this pattern by plotting the fitted nonparametrically smoothed relationship between persistence to 10<sup>th</sup> grade and mathematics score for low-income urban students, indicating that the probability of persisting jumps at the cut score between Passing and Failing.

TABLE 3 ABOUT HERE

---

<sup>22</sup> In the bottom panel, we use predicted math score because of the IV approach used for this analysis.

FIGURE 5 ABOUT HERE

(4) Effect of failing the English language arts examination on high school graduation<sup>23</sup>

Inspecting raw data in Massachusetts suggests that the mathematics examination is a larger hurdle to on-time graduation than the ELA examination. Most students who failed the 10<sup>th</sup> grade ELA examination also failed the mathematics test, while among students who only failed one of the tests, three times as many failed mathematics as ELA. The ELA examination proves interesting, however, because detected patterns differ from the mathematics results. Failing the 10<sup>th</sup> grade ELA examination does not reduce the probability of graduation for low-income, urban students (or for another group of students) on the margin of passing. In Table 4, we present parameter estimates and approximate *p*-values from our local linear regression analyses, again using only observations that fall within our “optimal” window, centered on the cut score. We illustrate the relationship between ELA score and probability of graduation for low-income urban students in Figure 6. Here, the figure displays no discontinuous jump in the probability of graduating at the cut score, suggesting that failing the ELA examination does not affect students’ likelihood of on-time graduation.

TABLE 4 ABOUT HERE

FIGURE 6 ABOUT HERE

***V. Sensitivity Analyses***

As discussed above, for regression discontinuity analyses to identify a causal effect of failing the MCAS examinations on student graduation, several assumptions must hold. First, the

---

<sup>23</sup> Because the middle school ELA test for the 2006 cohort occurred in 7<sup>th</sup> grade, one year earlier than the mathematics test, the state data system, which began in 2001, cannot match students as accurately for this test. As a result, we cannot examine the effects that *Fear of Failing* the ELA examination may have on persistence to 10<sup>th</sup> grade.

rule that determines whether a student has passed or failed the examination must be exogenous and rigidly applied across all students, while all other observed and unobserved characteristics of the student must vary smoothly and continuously around the cut score. Second, the relationship linking the probability of graduation and test score must be estimated accurately in the immediate vicinity of the cut score. In this section, we address these two primary concerns and describe other sensitivity analyses that we conduct to assess the robustness of our results.

### Exogenous Establishment of Cut Scores

The cut scores established by the Massachusetts Department of Education serve as an extremely plausible source of exogenous variation and do indeed produce a sharp discontinuity in treatment. Because the raw score needed to pass the examination differs from year to year and is only calculated after students take examination, it seems highly unlikely that students could decide knowingly to fall just above, or just below, the cut score. Furthermore, the state DOE imposes these performance labels strictly, so that any student with a score of 20 points on the 2004 administration of the 10<sup>th</sup> grade mathematics examination failed, while any student with a score of 21 points passed. Thus, the discontinuity is both exogenous and sharp.

We performed several additional tests to verify the exogeneity of the MCAS cut score, as recommended by Imbens & Lemieux (2007). We examined a histogram of the 10<sup>th</sup> grade mathematics scores to explore continuity around the cut score. We find that 899 students just failed the exam, while 900 just passed it. We also examined histograms of other covariates not affected by the examination to identify any apparent discontinuities around the cut score and found none. Finally, we split our sample into students who passed and students who failed in order to estimate effects at “pseudo-discontinuities” declared at the median mathematics scores

of these subsamples. In all cases, we find no reasons to doubt the robustness of our findings.<sup>24</sup>

#### Accurate estimation of the relationship between graduation and MCAS mathematics score

For estimates of the treatment effect to be unbiased, we must predict credibly and precisely what the probability of graduation would have been for students who failed the MCAS mathematics examination if they had scored 21 points on the test. We address this issue by modeling the smooth relationship between the probability of graduation and test score nonparametrically, using a local linear regression approach. Here, our primary specification decision then involves the choice of bandwidth,  $h$ . Our preferred models use optimal bandwidths chosen through the cross-validation procedures described above.

To explore the sensitivity of our results to differences in bandwidth selection, we vary it systematically, refitting our principal smoothed nonparametric models in each case. In the top panel of Table 5, we present the fitted effects of failing the 10<sup>th</sup> grade mathematics examination on on-time graduation for each subgroup as a function of different bandwidths. In the middle and bottom panels of Table 5, we present parallel results for the effects of failing the 8<sup>th</sup> grade examination on persistence to 10<sup>th</sup> grade and for the effects of failing the 10<sup>th</sup> grade ELA examination on on-time high school graduation. Regardless of bandwidth, our main results are unchanged – for urban, low income students, failing the 8<sup>th</sup> grade mathematics examination reduces the probability of persisting to 10<sup>th</sup> grade and failing the 10<sup>th</sup> grade mathematics examination reduces the probability of on-time graduation. However, we find no effects for other groups of students or for any group failing the 10<sup>th</sup> grade ELA examination. Our estimates for the effect of failing the 8<sup>th</sup> grade examination for marginal urban students range from 2.7 to 3.7 percentage points, and are quite insensitive to bandwidth. Our estimates of the effect of failing

---

<sup>24</sup> The results of these analyses are available from the authors upon request

the 10<sup>th</sup> grade examination range from 5.8 to 13.1 percentage points. In all cases, we reject the null hypothesis that the parameter value is zero.

TABLE 5 ABOUT HERE

Finally, we explore the sensitivity of the results to the choice of functional form for the relationship between probability of on-time graduation and MCAS mathematics score. As an alternative to our smoothed nonparametric specification, we fit logistic regression models that incorporate only observations in selected narrow “windows” around the MCAS cutoff. The top panel of Table 6 contains the critical predicted logistic regression coefficients and standard errors from models in which we estimate the impact of failing the 10<sup>th</sup> grade mathematics examination on the probability of on-time graduation. To facilitate interpretation, the bottom panel contains estimates in probability units of the causal impact of failing on the fitted probability of on-time graduation for a typical student. The results from the logistic regression analysis mirror almost identically those provided by our nonparametric approach.

TABLE 6 ABOUT HERE

Definition of outcome variable

We choose to present our main analyses using on-time graduation as our primary outcome measure. However, one concern is that students who fail the MCAS may remain in school and graduate in subsequent years, or that they may drop out and earn a General Equivalency Diploma instead of graduating from high school. We find that our results are quite robust to the definition of our outcome. Here, we use three different outcome measures: graduated on-time or still enrolled in school; dropped out; graduated on-time or obtained a GED. As seen in Table 7, in all cases we find statistically significant effects of passing the examination ranging from 7.2 to 9.1 percentage points.

## TABLE 7 ABOUT HERE

### *VI. Discussion*

This paper addresses several important questions about the effects of the state accountability system on Massachusetts high school students. To put these effects in context, it is important to recall the evidence cited earlier. Under standards-based educational reforms, the average reading and mathematics performances of Massachusetts students have improved markedly. In 2007, the state's reading and mathematics performances on the NAEP ranked first in the nation. Thus, we do not see the evidence that we present as an attack on the demonstrably successful educational reform effort in Massachusetts. Instead, we document unanticipated consequences of efforts to prepare all students to meet the demands of 21<sup>st</sup> century life. These consequences are important and need to be at the center of efforts to make standards-based reforms work for all Massachusetts students in the years ahead.

To recap, we find that, for low-income urban students on the margin of passing, failing the 8<sup>th</sup> grade mathematics examination reduces the probability of persisting to 10<sup>th</sup> grade by three percentage points, while failing the 10<sup>th</sup> grade examination reduces the probability of on-time graduation by eight percentage points. We find no effects of failing for wealthier urban students or suburban students. Again, these estimates are only valid for students at the margins of passing the examination, under the high-stakes testing regime in Massachusetts.

Importantly, we know nothing about whether these students are better (or worse) off than they would have been in the absence of standards-based reform. However, low-income, urban students with essentially the same proficiency on the state test have substantially different graduation outcomes simply because they are categorized as “passing” or “failing” the examination. This effect raises an important challenge for urban school districts. We also have

no information about the extent to which the requirement to pass the MCAS affects the probability of on-time graduation for students well below the passing score. As a result, we cannot estimate how much of the state dropout rate for low-income urban youth is due to the imposition of the exit examination. However, because 60% of students who do not graduate on time actually pass the MCAS, failing the test is clearly only one of many factors that contribute to the dropout decision.

We see several complementary explanations for the finding that failing the 10<sup>th</sup> grade mathematics examination reduces the likelihood of graduation for urban students from low-income families, but not for more affluent or suburban students. First, we cannot distinguish whether just failing the examination causes these students to drop out or whether just passing it causes them to remain in school. Low-income urban students who pass may feel encouraged that they are doing well in school and may decide to persist to graduation. Similarly, students who pass may get more teacher attention or may be promoted more readily through school, leading to improved graduation outcomes.

On the other hand, low-income urban students who fail the examination may become discouraged or subject to institutional responses that reduce their likelihood of graduating on time. Families of low-income urban students may lack the resources to help them overcome the hurdle posed by failing the examination. Low-income urban students typically attend high schools in which many students have failed the 10<sup>th</sup> grade MCAS examinations. These schools are struggling to figure out how, with very limited resources, to respond to this problem. Finally, the interaction between school and home contexts may produce these effects. Interestingly, the different consequences for failing the ELA examination than for failing the mathematics examination suggest that urban schools may devote more resources to or be more successful at

remediation in reading and writing than in mathematics.

That suburban students, including those from low-income families, appear to face no barrier from failing the 10<sup>th</sup> grade MCAS mathematics test suggests that their schools have found ways to support both low-income and wealthier students who have failed. These suburban schools typically have many fewer students who fail the examination, so they can afford to provide more personalized attention and remediation. In some Massachusetts districts, schools match teachers with students who failed the exit examination in order to provide one-on-one tutoring. In such an environment, it is not surprising that these students may in fact have more in-school adult contact and encouragement than students who just passed, and may in fact graduate at greater rates.

That most students who fail the 10<sup>th</sup> grade mathematics examination retake it and that low-income urban students retake the test at similar rates as their wealthier urban or suburban peers are also encouraging. These findings suggest that these students are receiving the message that they should persist and retake the test. As a result, schools have time to work with these students and prepare them to meet the graduation requirements. However, low-income urban students are much less likely to pass this retest, even when comparing students with the same initial examination performance. Finding the explanation for this pattern is an important topic for research, with critical implications for improving equality of educational opportunity.

Our findings raise several questions for researchers, educators, and policymakers in Massachusetts and other states. First, the absence of effects of high-stakes testing on high school completion for suburban students (including those from low-income families) suggests that it is possible to overcome the initial disappointment associated with failing a high-stakes examination. Learning more about the initiatives that improve student retention could be helpful

for districts struggling to support many failing students. A related question that we intend to pursue in future work is whether some urban districts are more successful than others in supporting students who failed the 10<sup>th</sup> grade mathematics examination. If that is the case, then understanding the successful efforts of some urban districts might help others to improve their support to struggling mathematics learners.

Especially intriguing is the finding that marginally failing the high-stakes ELA examination does not reduce the probability that low-income urban 10<sup>th</sup> graders graduated on time, while marginally failing the mathematics examination does reduce the probability of on-time graduation. Why the difference? Do urban districts concentrate resources on programs to improve their low-income students' ELA skills? Does the structure of the examinations make remediation easier in ELA than in mathematics for students on the border of passing?

Our finding that the *Fear of Failing* the 10<sup>th</sup> grade examination induces some low-income urban students to drop out before even taking it raises additional questions. Failing the 8<sup>th</sup> grade examination gives students some sense of their probable performance on the 10<sup>th</sup> grade test, but discerning students should recognize that scores on either side of the cutoff are not substantively different. Nonetheless, we found a moderate effect of failing on persistence to 10<sup>th</sup> grade for these very students. What is the mechanism at play here? Does the “failing” label affect a student's self-concept? Do students pay attention only to the performance level that their score puts them in, not on how close they are to passing? Or, does this effect reflect school or parental responses, such as retaining students or removing them to private schools?

Another question concerns the extent to which the consequences of exit examinations depend on their content and format. The 10<sup>th</sup> grade MCAS mathematics test is relatively demanding compared to the exit examinations used by other states. Not only does it assess

students' skills in a range of topic areas, it does so with questions that contain relatively complex language. Also, some test items call for open-ended responses while others require students to explain their answers. Supporters of the Massachusetts examinations argue that good instruction in mathematics is the only way to prepare students to do well on the test, and that simply drilling students on released test items is not an effective way to improve MCAS scores. The payoff to drill, as opposed to good mathematics instruction, may vary among the examinations used by different states. This difference may influence the success of various remediation programs.

This research argues for the importance of examining heterogeneous effects. In future work, we hope to explore more fully the effects of failing on different groups of students, including those with limited English proficiency. It also raises the question of whether the types of differential impacts we observe in Massachusetts may also be present in other states, especially those that use relatively demanding exit examinations. A corollary is the importance of finding the explanations for any observed differential effects of exit examinations. Finding differences in the probability of retaking the examination between groups suggests one policy problem. Finding differences in success rates among those who do retake the examination, as we do, suggests a different problem. We need to understand more carefully what messages and remediation efforts low-income urban students are receiving that encourage them to retake the examination but do not prepare them for success. Finally, we wonder why the effect for urban students varies by income. Do wealthier students attend different schools, or do they receive additional support outside of school?

In summary, the requirement that high school students achieve passing scores on relatively rigorous state-administered examinations in order to obtain a high school diploma is a relatively new phenomenon in the United States. The content, format, and difficulty of such tests

vary widely across states, as do opportunities for re-taking the examinations and support for those who fail. Future research needs to go beyond the question of whether failing a particular exit examination affects the probability of high school graduation. It needs to examine the extent to which the consequences of failing an exit examination depend on the attributes of the examination, the testing system, the student, and the quality of support available to struggling students.

Figure 1. Comparison of recent Massachusetts and nationwide National Assessment of Educational Progress scaled scores, for 8<sup>th</sup> grade mathematics from 1992 to 2008.

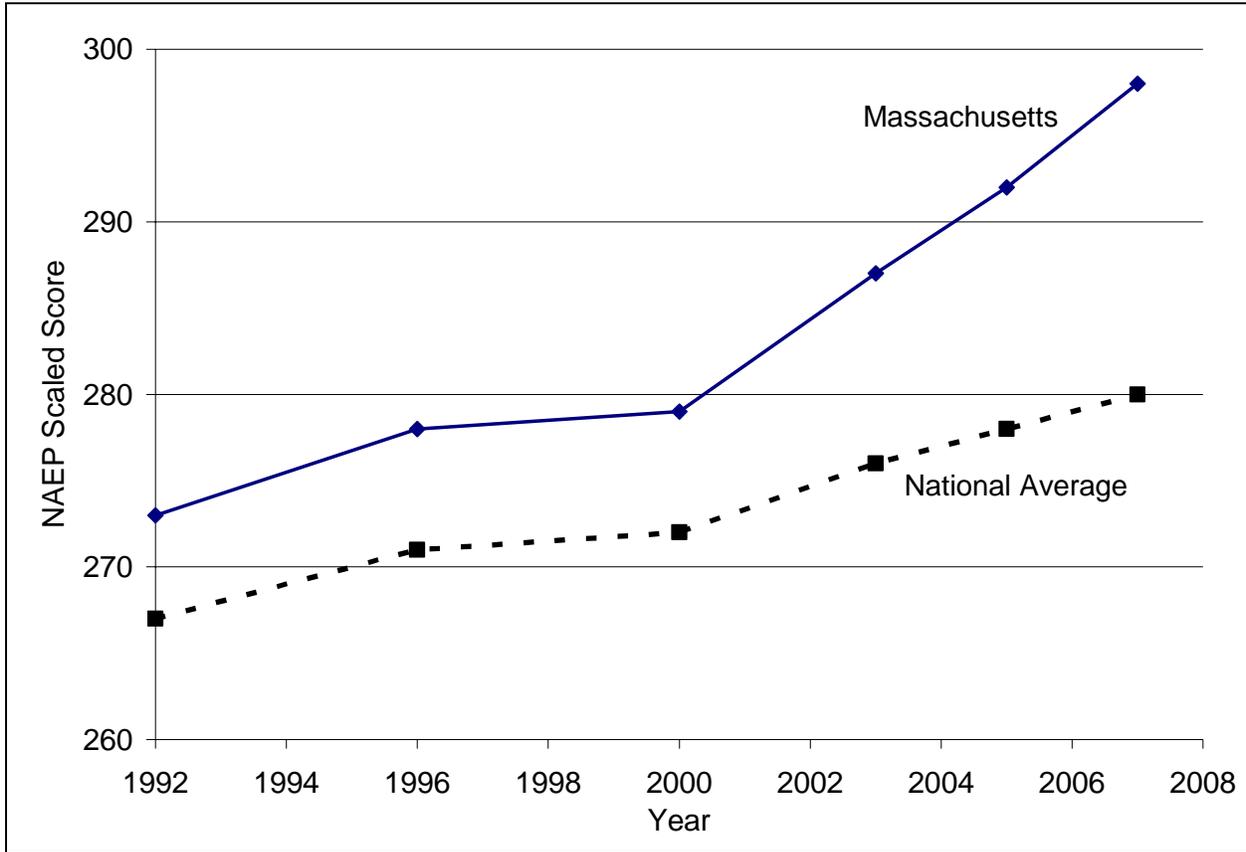
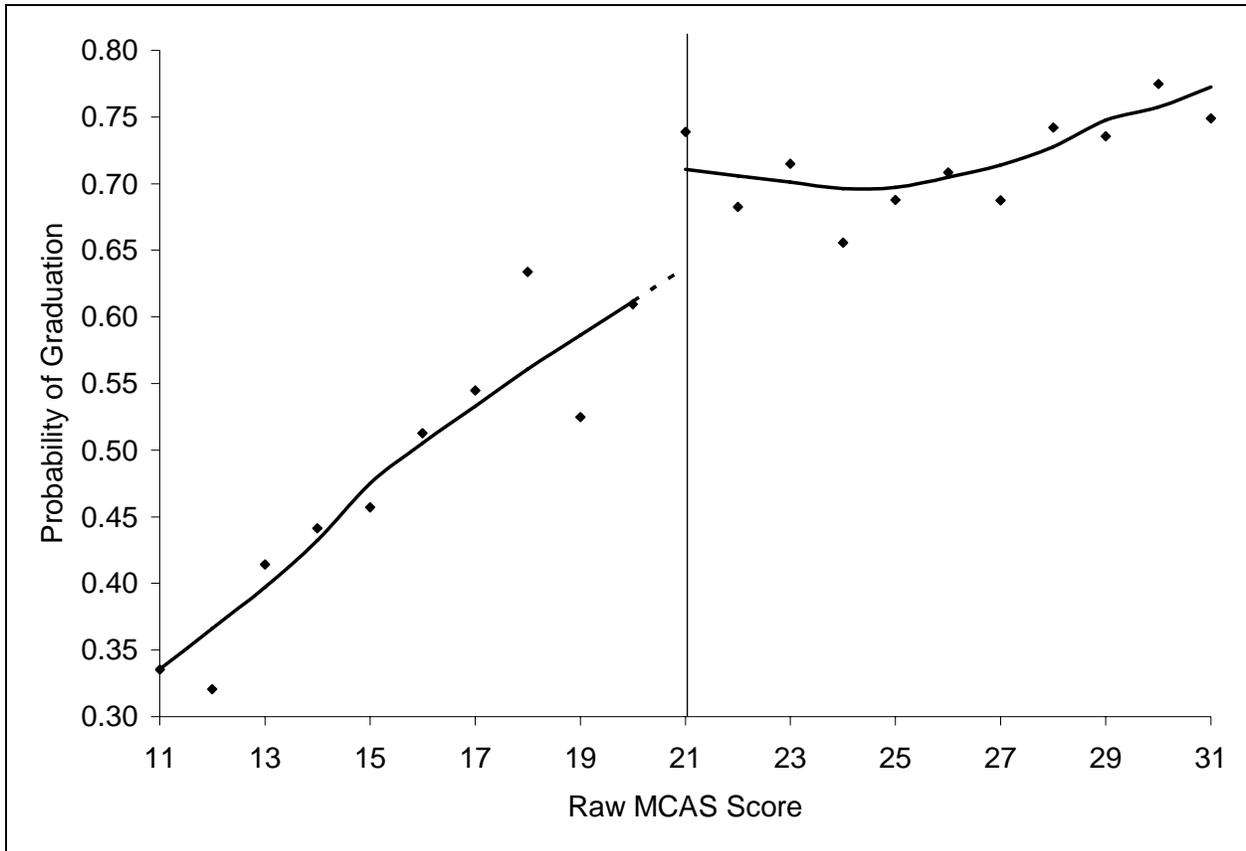


Figure 2. Fitted smoothed nonparametric relationship (bandwidth=6) between the probability of on-time graduation and 10<sup>th</sup> grade mathematics score for low-income urban students, with the sample mean probabilities of graduation overlaid.



We plot the nonparametric regression fit without student-level covariates.

Figure 3. Sample histogram presenting the frequencies of students who failed the 10<sup>th</sup> grade mathematics examination and who subsequently retook the examination, along with their performance on retest.

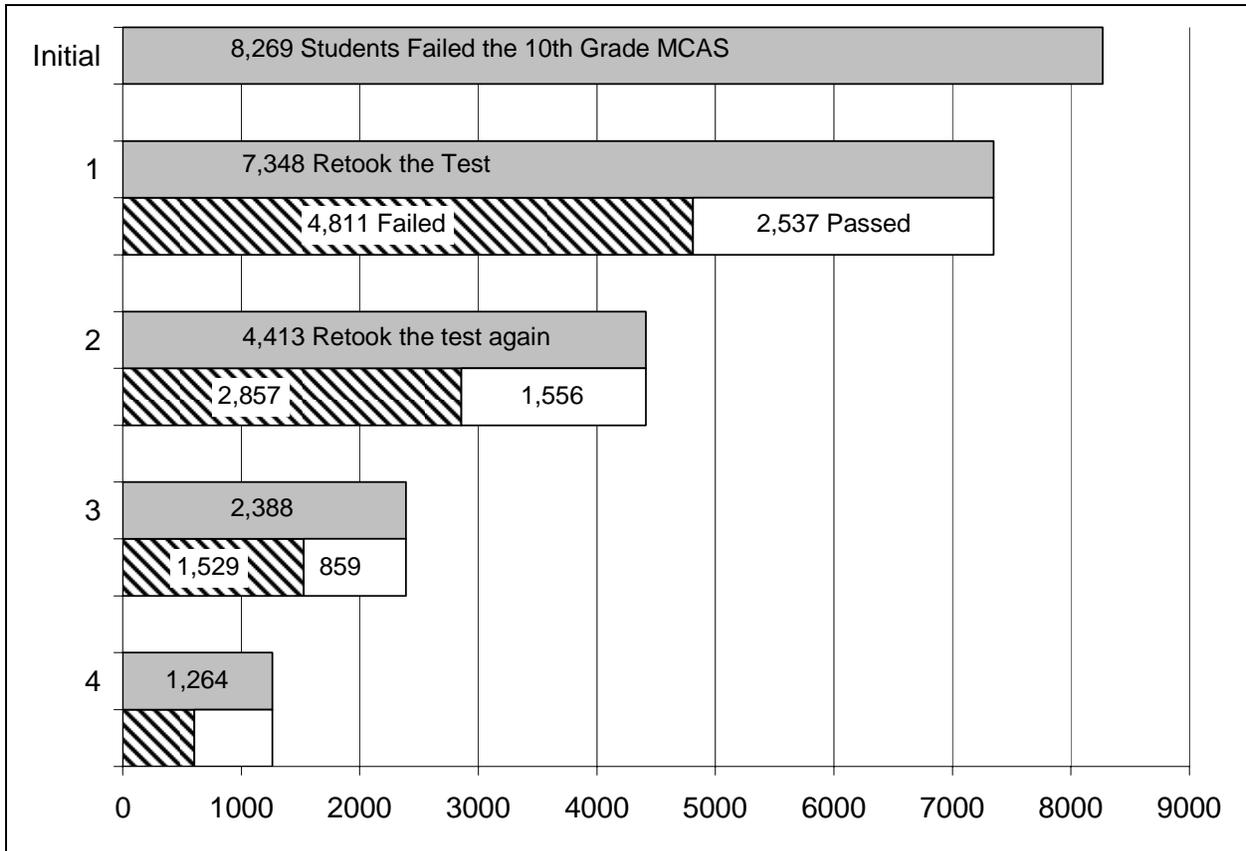


Figure 4. Fitted relationship (from Table 2) between the probability of retaking the examination (top panel) or passing the first retest (bottom panel) and initial 10<sup>th</sup> grade mathematics score for low-income urban students and suburban students who failed their first examination (plotted in the immediate region of the pass/fail cut-score for white female students not classified as special education or limited English proficient who just passed the ELA test) (n=8,225).

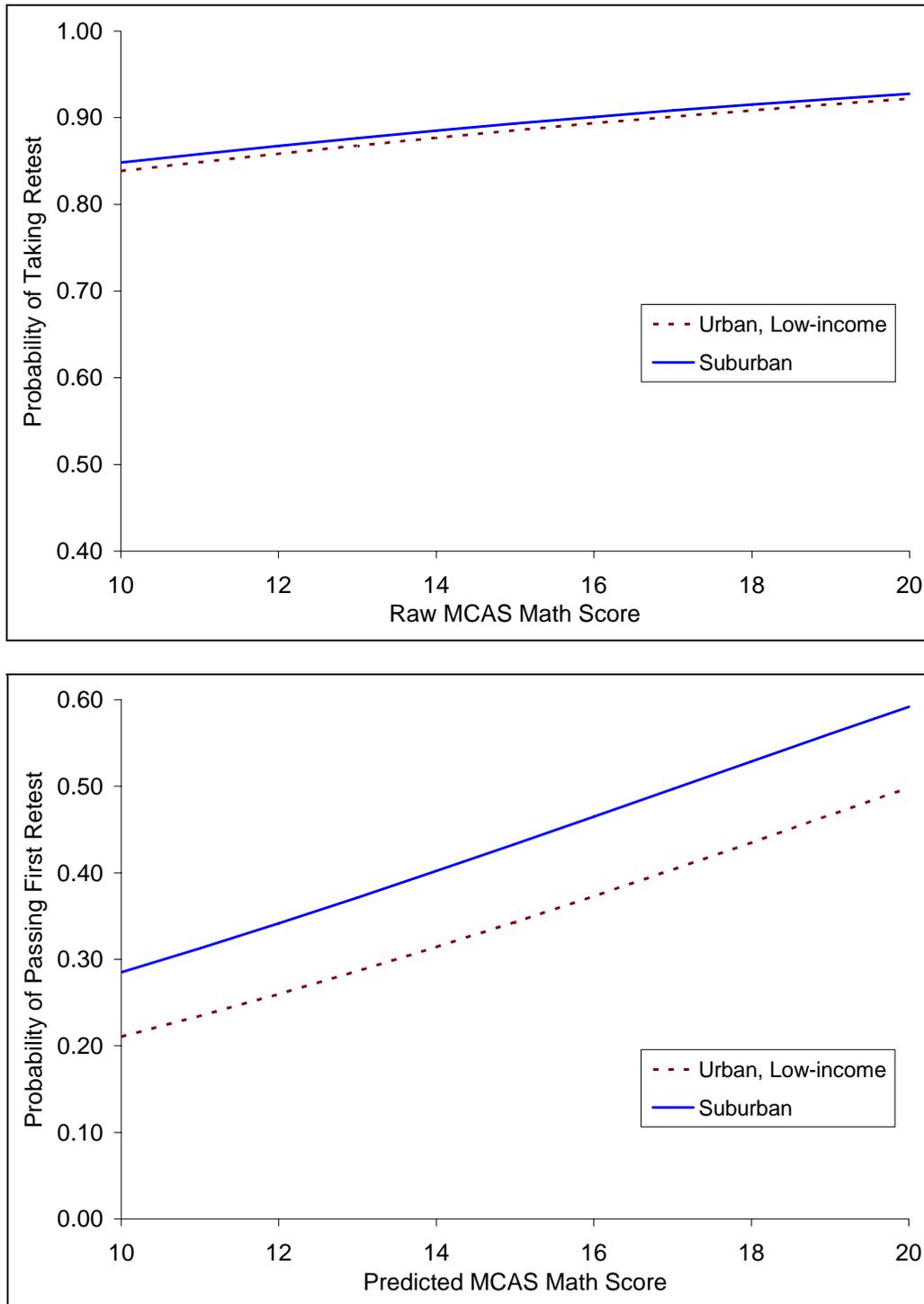
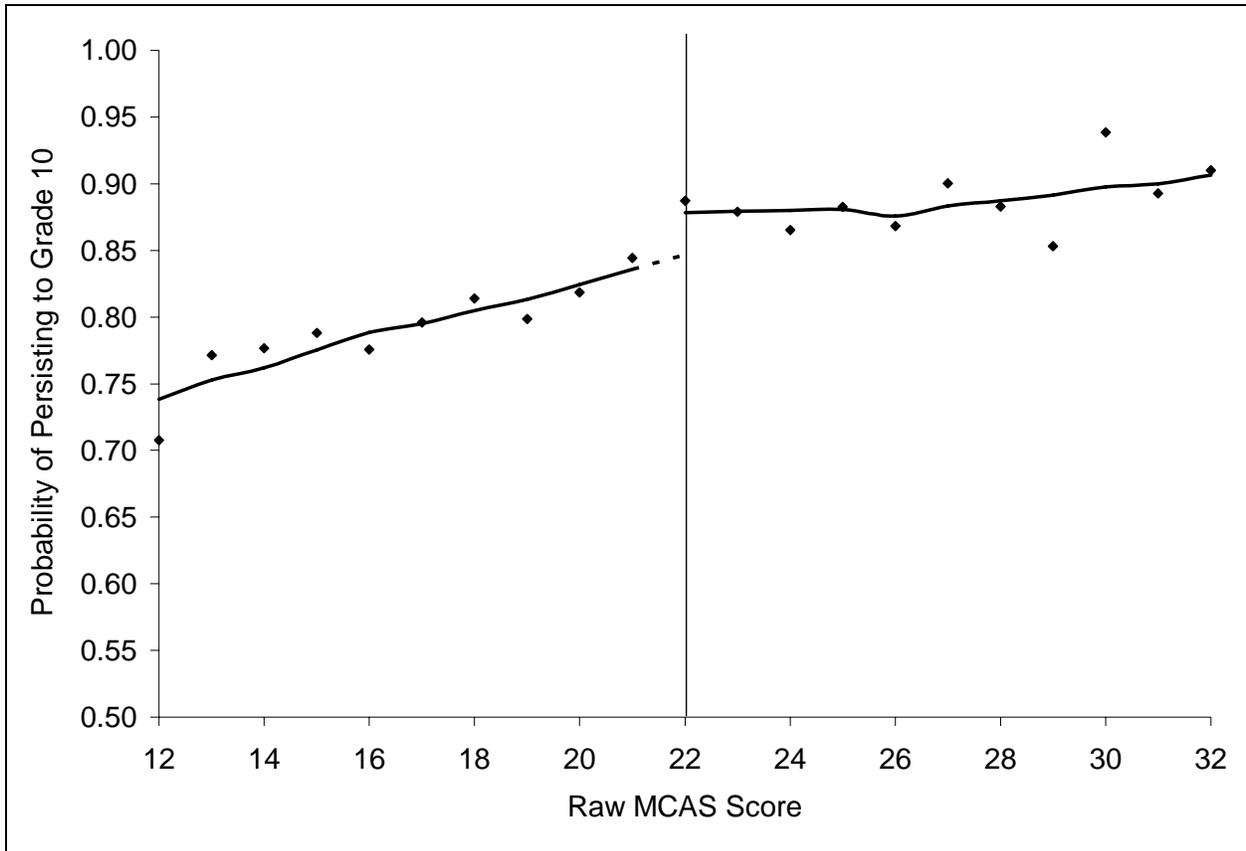
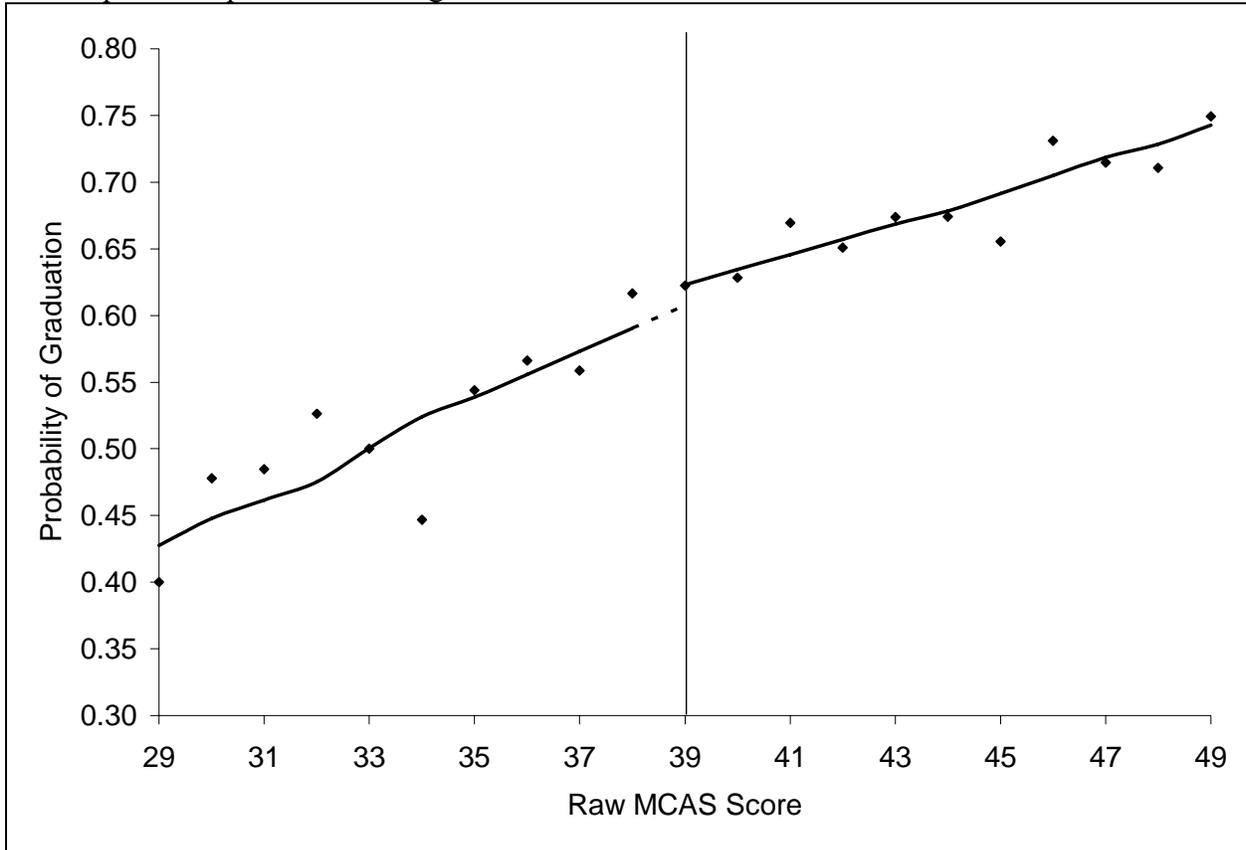


Figure 5. Fitted smoothed nonparametric relationship (bandwidth=6) between the probability of persisting to 10<sup>th</sup> grade and 8<sup>th</sup> grade mathematics score for low-income urban students, with the sample mean probabilities of graduation overlaid.



We plot the nonparametric regression fit without student-level covariates.

Figure 6. Fitted smoothed nonparametric relationship (bandwidth=8) between the probability of on-time high school graduation and 10<sup>th</sup> grade ELA score for low-income urban students, with the sample mean probabilities of graduation overlaid.



We plot the nonparametric regression fit without student-level covariates.

Table 1. Parameter estimates, standard errors, and approximate  $p$ -values at the cut score from the nonparametric regression analysis of the effect of failing the 10<sup>th</sup> grade mathematics examination on on-time graduation (from the single regression centered at the cut score with bandwidth  $h^*$ ).

Predictor	Urban Low-Income	Urban, Not Low-Income	Suburban, Low-Income	Suburban, Not Low-Income
Intercept	0.577 *** (0.033)	0.693 *** (0.049)	0.647 *** (0.039)	0.739 *** (0.025)
MATH	0.024 *** (0.007)	0.047 *** (0.012)	0.021 * (0.009)	0.025 *** (0.007)
PASS	0.080 * (0.033)	-0.052 (0.054)	0.023 (0.042)	-0.015 (0.027)
PASSxMATH	-0.031 *** (0.009)	-0.028 (0.014)	-0.022 (0.011)	-0.015 (0.008)
African-American	0.070 ** (0.022)	-0.003 (0.034)	0.103 ** (0.032)	0.041 (0.033)
Asian-American	0.059 (0.038)	0.126 (0.076)	0.163 ** (0.056)	0.028 (0.061)
Hispanic	-0.004 (0.022)	-0.091 * (0.040)	0.054 * (0.025)	-0.050 (0.033)
Mixed/Other Race	0.225 *** (0.068)	0.008 (0.110)	0.151 * (0.070)	-0.021 (0.090)
Native American	-0.068 (0.138)	-0.765 *** (0.027)	0.244 *** (0.059)	-0.017 (0.096)
Pacific Islander	-0.186 (0.261)	0.000	-0.195 (0.178)	-0.475 * (0.190)
Limited English Proficient	0.023 (0.024)	-0.103 (0.074)	-0.029 (0.043)	-0.085 (0.071)
Special Education	-0.015 (0.020)	0.031 (0.032)	0.031 (0.022)	0.050 *** (0.012)
Female	0.08 *** (0.016)	0.105 *** (0.025)	0.064 ** (0.020)	0.08 *** (0.012)
New Student	-0.057 * (0.024)	-0.058 (0.034)	-0.064 (0.033)	-0.081 *** (0.021)
R <sup>2</sup>	0.043	0.072	0.028	0.035
Bandwidth ( $h^*$ )	6	6	6	6
N	3469	1371	2172	4857

Notes: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

Table 2. Parameter estimates, standard errors, and approximate *p*-values from the probit instrumental variable regression analysis of the probability of retaking the examination and passing the first retest, among all students who originally failed (n=8,225).

Predictor	Probability of Retaking Test (Probit)	Probability of Passing Retest (IV Probit)
Intercept	1.609*** (0.066)	0.179* (0.071)
Urban*Lowinc	0.122 (0.065)	0.081 (0.054)
Urban	-0.163** (0.062)	-0.317*** (0.050)
MATH	0.043*** (0.004)	0.08*** (0.007)
ELA	0.025*** (0.003)	---
PASS (ELA)	-0.187** (0.065)	0.201*** (0.042)
African-American	0.362*** (0.065)	-0.118* (0.047)
Asian-American	0.049 (0.115)	0.062 (0.092)
Hispanic	0.071 (0.055)	-0.263*** (0.044)
Mixed/Other Race	0.854* (0.349)	-0.237 (0.163)
Native American	0.546 (0.383)	-0.057 (0.202)
Pacific Islander	1.148* (0.533)	-0.136 (0.246)
Limited English Proficient	0.195** (0.067)	-0.192*** (0.058)
Special Education	0.288*** (0.043)	-0.055 (0.033)
Female	0.08* (0.040)	-0.068* (0.030)
New Student	-0.353*** (0.049)	0.025 (0.043)
-2*Log Likelihood	4951	55922

Notes: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

Predictor	Probability of Retaking Test		Probability of Passing Retest	
	Probit	IV Probit	Probit	IV Probit
Intercept	1.609 *** (0.066)	1.938 *** (0.071)	-0.02 (0.051)	0.179 * (0.071)
Urban*Lowinc	0.122 (0.065)	0.122 (0.063)	0.078 (0.054)	0.081 (0.054)
Urban	-0.163 ** (0.062)	-0.138 * (0.060)	-0.329 *** (0.050)	-0.317 *** (0.050)
MATH	0.043 *** (0.004)	0.12 *** (0.006)	0.043 *** (0.004)	0.08 *** (0.007)
ELA	0.025 *** (0.003)	---	0.011 *** (0.002)	---
PASS (ELA)	-0.187 ** (0.065)	-0.04 (0.053)	0.137 ** (0.051)	0.201 *** (0.042)
African-American	0.362 *** (0.065)	0.312 *** (0.062)	-0.106 * (0.047)	-0.118 * (0.047)
Asian-American	0.049 (0.115)	0.049 (0.111)	0.061 (0.092)	0.062 (0.092)
Hispanic	0.071 (0.055)	0.049 (0.053)	-0.257 *** (0.045)	-0.263 *** (0.044)
Mixed/Other Race	0.854 * (0.349)	0.798 * (0.332)	-0.235 (0.164)	-0.237 (0.163)
Native American	0.546 (0.383)	0.479 (0.366)	-0.04 (0.203)	-0.057 (0.202)
Pacific Islander	1.148 * (0.533)	1.195 * (0.507)	-0.19 (0.248)	-0.136 (0.246)
Limited English Proficient	0.195 ** (0.067)	0.101 (0.064)	-0.155 ** (0.059)	-0.192 *** (0.058)
Special Education	0.288 *** (0.043)	0.281 *** (0.042)	-0.059 (0.033)	-0.055 (0.033)
Female	0.08 * (0.040)	0.1 ** (0.039)	-0.081 ** (0.031)	-0.068 * (0.030)
New Student	-0.353 *** (0.049)	-0.318 *** (0.048)	0.018 (0.043)	0.025 (0.043)
-2*Log Likelihood	4951	9291	51582	55922

Notes: \* p<0.05, \*\* p<0.01, \*\*\* p<0.001.

Table 3. Parameter estimates, standard errors, and approximate  $p$ -values at the cut score from the nonparametric regression analysis of the effect of failing the 8<sup>th</sup> grade mathematics examination on persistence to 10<sup>th</sup> grade (from the single regression centered at the cut score with bandwidth  $h^*$ ).

Predictor	Urban Low-Income	Urban, Not Low-Income	Suburban, Low-Income	Suburban, Not Low-Income
Intercept	0.796 *** (0.019)	0.815 *** (0.025)	0.886 *** (0.018)	0.940 *** (0.008)
MATH (8 <sup>th</sup> Grade)	0.012 ** (0.004)	-0.001 (0.006)	0.008 (0.005)	0.006 ** (0.002)
PASS (8 <sup>th</sup> Grade)	0.027 (0.019)	0.048 (0.028)	-0.002 (0.020)	-0.008 (0.009)
PASSxMATH	-0.010 (0.005)	0.001 (0.008)	-0.003 (0.005)	-0.002 (0.002)
African-American	0.078 *** (0.013)	-0.080 *** (0.023)	0.026 (0.017)	-0.030 (0.017)
Asian-American	0.049 ** (0.019)	-0.028 (0.047)	0.030 (0.028)	-0.035 (0.024)
Hispanic	0.028 * (0.013)	-0.051 * (0.024)	0.014 (0.014)	-0.054 ** (0.018)
Mixed/Other Race	0.189 *** (0.011)	0.110 *** (0.030)	0.111 *** (0.008)	0.049 *** (0.012)
Native American	-0.145 (0.110)	0.032 (0.114)	0.098 *** (0.008)	-0.040 (0.068)
Pacific Islander	-0.021 (0.132)	0.138 *** (0.018)	-0.060 (0.117)	0.063 *** (0.005)
Limited English Proficient	0.007 (0.017)	-0.223 *** (0.066)	-0.055 (0.040)	-0.082 (0.064)
Special Education	0.000 (0.015)	-0.009 (0.021)	0.010 (0.012)	0.005 (0.005)
Female	0.029 ** (0.010)	0.035 * (0.014)	0.023 * (0.010)	0.011 * (0.004)
R <sup>2</sup>	0.023	0.025	0.011	0.006
Bandwidth ( $h^*$ )	6	6	6	6
N	5709	2828	3759	13160

Notes: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

Table 4. Parameter estimates, standard errors, and approximate  $p$ -values at the cut score from the nonparametric regression analysis of the effect of failing the 10<sup>th</sup> grade ELA examination on on-time graduation (from the single regression centered at the cut score with bandwidth  $h^*$ ).

Predictor	Urban Low-Income	Urban, Not Low-Income	Suburban, Low-Income	Suburban, Not Low-Income
Intercept	0.519 *** (0.034)	0.617 *** (0.059)	0.641 *** (0.042)	0.716 *** (0.031)
ELA	0.018 ** (0.006)	0.035 ** (0.013)	0.022 * (0.009)	0.021 ** (0.007)
PASS	0.011 (0.034)	-0.052 (0.067)	-0.016 (0.044)	0.022 (0.033)
PASSxELA	-0.005 (0.007)	-0.009 (0.014)	-0.015 (0.010)	-0.011 (0.008)
African-American	0.059 ** (0.022)	-0.020 (0.037)	0.107 *** (0.030)	0.069 * (0.033)
Asian-American	0.123 *** (0.030)	0.033 (0.065)	0.205 *** (0.037)	0.100 * (0.045)
Hispanic	0.015 (0.022)	-0.091 * (0.042)	0.040 (0.025)	-0.059 (0.036)
Mixed/Other Race	0.121 (0.079)	-0.058 (0.139)	0.204 *** (0.056)	0.168 *** (0.046)
Native American	0.059 (0.152)	0.306 *** (0.077)	0.356 *** (0.031)	0.088 (0.075)
Pacific Islander	-0.183 (0.199)	---	-0.402 (0.225)	-0.244 (0.176)
Limited English Proficient	0.102 *** (0.021)	-0.098 (0.069)	0.080 * (0.036)	-0.004 (0.063)
Special Education	0.021 (0.019)	0.021 (0.033)	0.042 * (0.021)	0.029 * (0.013)
Female	0.049 ** (0.016)	0.080 ** (0.027)	0.058 ** (0.019)	0.055 *** (0.012)
New Student	0.004 (0.018)	-0.031 (0.032)	-0.064 * (0.025)	-0.124 *** (0.018)
R <sup>2</sup>	0.039	0.081	0.040	0.046
Bandwidth ( $h^*$ )	8	8	8	8
N	3820	1180	2281	4449

Notes: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

Table 5. Estimated causal impacts of failing the 10<sup>th</sup> grade mathematics, 8<sup>th</sup> grade mathematics, and 10<sup>th</sup> grade ELA examinations, for different bandwidths by subgroup, with standard errors in parentheses. Results for the optimal bandwidth,  $h^*$ , appear in bold.

<b>Panel I: 10<sup>th</sup> Grade Mathematics</b>					
<b>Group</b>	<b>Bandwidth (h)</b>				
	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>
Urban, Low Income	0.131 ** (0.041)	0.103 ** (0.036)	<b>0.080 *</b> <b>(0.033)</b>	0.065 * (0.031)	0.058 * (0.029)
Urban, Not Low Income	-0.025 (0.067)	-0.024 (0.059)	<b>-0.052</b> <b>(0.054)</b>	-0.007 (0.050)	-0.007 (0.046)
Suburban, Low Income	-0.050 (0.052)	-0.022 (0.045)	<b>0.023</b> <b>(0.042)</b>	0.009 (0.038)	0.003 (0.036)
Suburban, Not Low Income	0.000 (0.034)	-0.018 (0.030)	<b>-0.015</b> <b>(0.027)</b>	-0.016 (0.025)	-0.027 (0.024)

<b>Panel II: 8<sup>th</sup> Grade Mathematics</b>					
<b>Group</b>	<b>Bandwidth (h)</b>				
	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>
Urban, Low Income	0.032 (0.024)	0.029 (0.021)	<b>0.027</b> <b>(0.019)</b>	0.037 * (0.018)	0.034 * (0.017)
Urban, Not Low Income	0.047 (0.036)	0.027 (0.031)	<b>0.048</b> <b>(0.028)</b>	0.029 (0.026)	0.013 (0.025)
Suburban, Low Income	0.011 (0.024)	0.012 (0.022)	<b>-0.002</b> <b>(0.020)</b>	0.008 (0.018)	0.014 (0.017)
Suburban, Not Low Income	-0.008 (0.011)	-0.003 (0.010)	<b>-0.008</b> <b>(0.009)</b>	-0.010 (0.008)	-0.013 (0.008)

<b>Panel III: 10<sup>th</sup> Grade ELA</b>					
<b>Group</b>	<b>Bandwidth (h)</b>				
	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
Urban, Low Income	-0.002 (0.039)	0.006 (0.036)	<b>0.011</b> <b>(0.034)</b>	0.019 (0.032)	0.010 (0.030)
Urban, Not Low Income	-0.121 (0.076)	-0.090 (0.072)	<b>-0.052</b> <b>(0.067)</b>	-0.023 (0.063)	0.023 (0.060)
Suburban, Low Income	-0.046 (0.052)	-0.031 (0.048)	<b>-0.016</b> <b>(0.044)</b>	0.003 (0.042)	0.006 (0.007)
Suburban, Not Low Income	-0.017 (0.038)	-0.001 (0.035)	<b>0.022</b> <b>(0.033)</b>	0.027 (0.031)	0.032 (0.030)

Notes: \* p<0.05, \*\* p<0.01, \*\*\* p<0.001.

Table 6. Estimated causal impact of failing the 10<sup>th</sup> grade mathematics examination on on-time high school graduation from a logistic regression model, for samples within windows of different widths around the cut score. Panel I presents the estimated logistic regression coefficients, with standard errors in parentheses; Panel II presents the fitted differences in the probability of graduation for a typical student. Results for the optimal bandwidth,  $h^*$ , appear in bold

<b>Panel I: Logistic regression coefficients (standard errors in parentheses)</b>					
<b>Group</b>	<b>Width of window around discontinuity</b>				
	<b>+/- 4</b>	<b>+/- 5</b>	<b>+/- 6</b>	<b>+/- 7</b>	<b>+/- 8</b>
Urban, Low Income	0.596 ** (0.182)	0.473 ** (0.160)	<b>0.378 *</b> <b>(0.147)</b>	0.307 * (0.135)	0.277 * (0.127)
Urban, Not Low Income	-0.111 (0.308)	-0.113 (0.269)	<b>-0.231</b> <b>(0.249)</b>	-0.024 (0.229)	-0.028 (0.213)
Suburban, Low Income	-0.222 (0.253)	-0.084 (0.221)	<b>0.133</b> <b>(0.200)</b>	0.061 (0.183)	0.030 (0.171)
Suburban, Not Low Income	0.002 (0.186)	-0.101 (0.166)	<b>-0.062</b> <b>(0.150)</b>	-0.066 (0.140)	-0.114 (0.131)

<b>Panel II: Probability of graduation</b>					
<b>Group</b>	<b>Width of window around discontinuity</b>				
	<b>+/- 4</b>	<b>+/- 5</b>	<b>+/- 6</b>	<b>+/- 7</b>	<b>+/- 8</b>
Urban, Low Income	0.127 **	0.101 **	<b>0.079 *</b>	0.065 *	0.058 *
Urban, Not Low Income	-0.025	-0.022	<b>-0.041</b>	-0.005	-0.005
Suburban, Low Income	-0.041	-0.016	<b>0.027</b>	0.012	0.006
Suburban, Not Low Income	0.000	-0.016	<b>-0.010</b>	-0.010	-0.017

The “typical” student in the regression discontinuity sample is a white female, not classified as either LEP or as special education.

Table 7. Parameter estimates, standard errors, and approximate  $p$ -values at the cut score from the nonparametric regression analysis of the effect of failing the 10<sup>th</sup> grade mathematics examination on three different graduation outcomes, for urban, low-income students (from the single regression centered at the cut score with bandwidth  $h^*$ ).

Predictor	Graduated or still enrolled	Dropped Out	Graduated or earned GED
Intercept	0.660 *** (0.031)	0.273 *** (0.027)	0.607 *** (0.033)
MATH	0.015 * (0.007)	0.002 (0.006)	0.025 *** (0.007)
PASS	0.091 ** (0.030)	-0.072 ** (0.026)	0.081 * (0.033)
PASSxMATH	-0.024 ** (0.008)	0.001 (0.007)	-0.031 *** (0.009)
African-American	0.117 *** (0.020)	-0.089 *** (0.017)	0.059 ** (0.022)
Asian-American	0.081 * (0.035)	-0.067 * (0.030)	0.061 (0.038)
Hispanic	0.039 (0.021)	-0.054 ** (0.018)	-0.011 (0.021)
Mixed/Other Race	0.241 *** (0.054)	-0.152 *** (0.044)	0.204 ** (0.068)
Native American	-0.081 (0.135)	-0.139 (0.072)	-0.086 (0.138)
Pacific Islander	0.145 (0.201)	-0.015 (0.189)	-0.208 (0.261)
Limited English Proficient	0.027 (0.022)	-0.020 (0.018)	0.007 (0.024)
Special Education	-0.003 (0.019)	-0.012 (0.016)	-0.028 (0.020)
Female	0.031 * (0.015)	-0.039 ** (0.012)	0.073 *** (0.016)
New Student	-0.071 ** (0.023)	0.003 (0.018)	-0.050 * (0.024)
R <sup>2</sup>	0.036	0.018	0.045
Bandwidth ( $h^*$ )	6	6	6
N	3469	3469	3469

Notes: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

## Appendix A

To explore the extent to which measurement error may affect OLS estimates of retest success by subgroup and whether IV estimates remove this source of bias, we ran a simulation designed to mirror our analyses. Here, we simulated mathematics, ELA, and mathematics retest scores for two groups of students: suburban and urban, low income. Using sample data and published reliabilities, we derived estimates of the true score means, true score covariance matrices, and error variances on each of the tests.

Using these estimates, we drew true scores on both the “math” and “ELA” tests for 10,630 ULI students and 49,378 suburban students from a multivariate normal distribution with the sample true covariance matrix. We then drew three sets of mean zero errors – two for math and one for ELA, using the appropriate error variances. We added the appropriate errors to the true scores to obtain mathematics “test” and “retest” scores and ELA “test” scores for each observation. We iterated this process, drawing 10,000 different samples.

Our simulation results confirm that truncating the sample to include only students who fail does produce mechanical differences in retest success. We find that the average urban, low-income students who fails the first mathematics test scores 3 points lower on their retest than suburban students who fail. Comparing students with the same initial test scores near the cutoff, low-income urban students are six to seven percentage points less likely to pass the retest than similarly able suburban students. OLS regression reveals a statistically significant relationship between retest score and urban, low-income status in 99.8% of cases. Again, these differences arose mechanically, without any changes in the underlying true score distribution.

Hanushek & Rivkin (2006) suggest that using another test score to instrument for the initial mathematics test can resolve this problem. Here, we use the “ELA” test as an instrument for the mathematics test. We implement this approach with two-stage least squares in each of the 10,000 datasets constructed above. We find a statistically significant relationship (with  $\alpha=0.05$ ) between retest score and urban, low-income status in just 4.8% of the samples, within the tolerance that we could expect by chance. Thus, the simulation appears to confirm that the IV approach resolves this issue.

## References

- Carnoy, M. (2005). Have state accountability and high-stakes tests influenced student progression rates in high school? *Educational Measurement: Issues and Practice*. Winter, 19-31.
- Carnoy, M. & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.
- Center on Education Policy. (2007). "It's different now": *How exit examinations are affecting teaching and learning in Jackson and Austin*. Retrieved June 26, 2008, from <http://www.cep-dc.org/highschool/exit/JacksonAustin/Jackson&Austin.pdf>
- Clarke, M., Haney, W. & Madaus, G. (2000). High stakes testing and high school completion. *National Board on Educational Testing and Public Policy Statement*, 1(3). Retrieved June 26, 2008, from <http://www.bc.edu/research/nbetpp/publications/v1n3.html>.
- Cornell, D.G., Krosnik, J.A. & Chang, L. (2006). Student reactions to being wrongly informed of failing a high-stakes test: The case of the Minnesota Basic Standards test. *Educational Policy*, 20(5), 718-751.
- Dee, T.S. & Jacob, B.A. (2006). Do high school exit exams influence educational attainment or labor market performance? Cambridge, MA: NBER Working Paper 12199. Retrieved June 26, 2008, from <http://www.nber.org/papers/w12199>.
- Evers, W.M., & Walberg, H.J., eds. (2002). *School accountability*. Stanford, CA: Hoover Institution Press.
- Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420): 998-1004.

- Finn, C.E., Julian, L., & Petrilli, M.J. (2006). *The state of state standards*. Washington, D.C.: The Fordham Foundation. Retrieved March 26, 2008 from <http://www.edexcellence.net/foundation/publication/publication.cfm?id=358>.
- Greene, J.P. & Winters, M.A. (2004). *Education working paper: Pushed out or pulled up? Exit examinations and dropout rates in public high schools*. New York: Center for Civic Innovation at the Manhattan Institute for Policy Research.
- Griffin, B.W. & Heidorn, M.H. (1996). An examination of the relationship between minimum competency test performance and dropping out of high school. *Educational Evaluation and Policy Analysis*, 18(3), 243-252.
- Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1): 201-209.
- Hanushek, E.A., & Rivkin, S.G. (2006). School quality and the black-white achievement gap. Cambridge, MA: NBER Working Paper 12651. Retrieved June 26, 2008, from <http://www.nber.org/papers/w12651>.
- Heubert, J.P. & Hauser, R.M., eds. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press, 1999.
- Imbens, G. & Lemieux, T. (2007). Regression discontinuity designs: A guide to practice. Cambridge, MA: NBER Working Paper 13039. Retrieved June 26, 2008, from <http://www.nber.org/papers/w13039>.
- Jacob, B.A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, 23(2), 99-121.
- Jones, M.G., Jones, B.D., & Hargrove, T.Y. (2003). *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

- Levy, F. & Murnane, R. (2004). *The new division of labor: How computers are creating the next job market*. Princeton, N.J.: Princeton University Press; New York: Russell Sage Foundation.
- Lochner, L., & Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review*, 94(1), 155-189.
- Lleras-Muney, A. (2004). The relationship between education and adult mortality in the United States. *Review of Economic Studies*, 72(1).
- Ludwig, J. and Miller D. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, 122(1), 159-208.
- Martorell, F. (2005). Does failing a high school graduation exam matter? Unpublished working paper: Author.
- Massachusetts Department of Education. (2002). *2001 MCAS technical report*. Retrieved June 26, 2008, from <http://www.doe.mass.edu/mcas/2002/news/01techrpt.pdf>.
- Massachusetts Department of Education. (2005). *2004 MCAS technical report*. Retrieved June 26, 2008, from <http://www.doe.mass.edu/mcas/2005/news/04techrpt.pdf>.
- Massachusetts Department of Education. (2007). *2007 NAEP tests: Summary of results for Massachusetts*. Retrieved March 26, 2008, from <http://www.doe.mass.edu/news/news.asp?id=3692>.
- National Center for Education Statistics. (2008). State comparisons: National Assessment of Educational Progress (NAEP). Washington, DC: U.S. Department of Education. Retrieved April 5, 2008 from <http://nces.ed.gov/nationsreportcard/nde/statecomp/>

- Neal, D. & Schanzenbach, D.W. (2007). Left behind by design: Proficiency counts and test-based accountability. Unpublished working paper: Author.
- Nichols, S.L., Glass, G.V, & Berliner, D.C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives*, 14(1). Retrieved June 26, 2008, from <http://epaa.asu.edu/epaa/v14n1/>.
- Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health, and happiness from compulsory schooling. *Journal of Political Economy*, 91(2007), 2213-2229.
- Quality Counts. (2006). Quality counts at 10: A decade of standards-based education. *Education Week*, 25(17): 74.
- Reardon, S.F. & Galindo, C. (April, 2002). *Do high-stakes tests affect students' decisions to drop out of school? Evidence from NELS*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.
- Thistlethwaite, D.L. & Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the *ex post facto* experiment. *The Journal of Education Psychology*, 51(6), 309-317.
- Thomas, R.M. (2005). *High stakes testing: Coping with collateral damage*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Warren, J.W., Jenkins, K.N., & Kulick, R.B. (2006). High school exit examinations and state-level completion and GED rates, 1975 through 2002. *Educational Evaluation and Policy Analysis*, 28:2, 131-152.