The Impact of Test Score Labels on Human-Capital Investment Decisions

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

ABSTRACT

Students receive abundant information about their educational performance, but how this information affects future educational-investment decisions is not well understood. Increasingly, results from state-mandated standardized tests are an important source of information. Students receive a score and a label that summarizes their performance on these tests. Using a regression-discontinuity design, we find persistent effects of earning a more positive label on the college-going decisions of urban, low-income students. These findings are important not only for understanding students' educational-investment decisions and the consequences of state testing practices but also for researchers using the regression-discontinuity design.

I. Introduction

Over the past few decades, states have implemented test-based accountability programs to monitor student progress toward mastering content standards. Under the No Child Left Behind Act, states must test all students in both mathematics and English language arts (ELA) in Grades 3–8 and on a single occasion in high school.

John P. Papay is an assistant professor of education and economics at Brown University. Richard J. Murnane is the Thompson Research Professor at the Harvard Graduate School of Education and a research associate at the National Bureau of Economic Research. John B. Willett is the Eliot Research Professor at the Harvard Graduate School of Education. *The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E100013 to Harvard University and by the Spencer Foundation. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the Spencer Foundation. The authors thank Carrie Conaway, the Associate Commissioner of Planning, Research, and Delivery Systems for the Massachusetts Department of Elementary and Secondary Education (DESE) for providing the data and for answering many questions about data collection procedures. Scholars may acquire the data by requesting access from the Massachusetts DESE. The authors are willing to advise other scholars. Address correspondence to John Papay (john_papay@brown.edu).*

[Submitted July 2014; accepted November 2014]

ISSN 0022-166X E-ISSN 1548-8004 © 2016 by the Board of Regents of the University of Wisconsin System

doi:10.3368/jhr.51.2.0713-5837R

Testing has now become a core enterprise in the educational system, which spends an estimated $1.7 billion annually on this activity (Chingos 2012).

One key feature of test-based accountability systems in the United States is that every student receives not only a test score but also a label based on their performance. Massachusetts, the state that we study, assigns students labels of *Warning/ Failing*, *Needs Improvement*, *Proficient*, or *Advanced* by determining cut-points with which it divides the finer-grained test score distribution into performance regions. In this paper, we use a regression-discontinuity design to examine the effect of these labels on students' educational investment decisions, including their decision to enroll in college. These labels provide no additional information beyond the test scores on which they are based; they are simply coarse summaries of a student's performance. We focus on responses to labels that have no state-defined consequences for students.

We find that these test score labels do influence students' decisions about post-secondary education. In particular, we find that the effects are concentrated in mathematics among a particular group of students—the 18 percent of students in Massachusetts who have low family income and attend urban schools. We find no effects for suburban students or students from higher-income families or for any students on the English language arts (ELA) exam. We find that being labeled *Advanced* rather than *Proficient* on the tenth grade mathematics test increases by five percentage points the probability that urban, low-income students at the margin will attend college. This is a substantial effect, given that only 65 percent of such students who barely score *Advanced* attend college. We find particularly strong effects for an important subgroup of urban, low-income students—those who do not plan to attend a four-year college; this group contains approximately 5–8 percent of students in the state.

These results demonstrate that new performance information students receive, even during high school, can affect their educational futures. That dividing a continuous performance distribution into discrete categories affects students' post-secondary educational enrollments is clearly an important, unintended consequence of state testing policies as they have been implemented. These findings are important not only for policymakers designing test-based accountability systems but also for researchers using the regression-discontinuity design. In recent years, economists have exploited test-based cutoffs that assign students to different treatments in order to draw causal conclusions about these treatments. To the extent that students respond to any associated performance labels, assignment to treatment may be confounded with an effect of labeling, producing biased estimates of treatment effects.

In the next section, we review briefly the ways in which performance labels may influence students' subsequent educational decisions. We then describe our data sources, key measures, and data-analytic strategy. In Sections IV and V, we present our main findings and describe sensitivity analyses that we conduct to assess the robustness of our results. Finally, we conclude with a discussion of our findings and their implications for educational practice in a regime of test-based accountability and for research using the regression-discontinuity design.

II. Background and Context

A. Standards-Based Reform and State Testing

In 2001, as part of No Child Left Behind (the reauthorization of the Elementary and Secondary Education Act (ESEA)), the federal government required that all states taking federal ESEA funds adopt academic standards, develop an annual testing program to assess student progress toward those standards, and define what proficient mastery of those standards meant. This legislation set the goal that all American public school students should be proficient in mathematics and English language arts (ELA) by 2014. It also mandated that all schools must demonstrate annually that they are making "Adequate Yearly Progress" (AYP) toward this goal for students in a variety of demographic subgroups including racial minorities and students with special educational needs or limited English proficiency. Schools that fail to meet AYP for several years in a row are subject to increasingly severe sanctions.

Here, we focus on the information that students receive about their test performance. In Massachusetts, each student receives a detailed report several months after taking the test that includes their test score, a confidence interval around the score, and a performance label. In the appendix, we include an example of one such report. The reports also contain interpretive information (not shown) to help students and parents make sense of their test scores. Thus, students receive a substantial amount of information about their test performance in addition to the score and label. Although the label adds no additional information to the test score, it remains the easiest and most intuitive element of the report to interpret.

B. Educational Performance Information and Human Capital Investment

There are clearly many determinants of educational investment decisions, and the ways in which adolescents make them are not well understood (see Murnane 2013 for a recent overview). According to standard economic models of educational investment, an individual's abilities, or at least his or her perceptions of those abilities, are important determinants of the benefits and costs of further education. Although these returns and costs may be difficult for students to assess, their educational performance is not. Throughout their school careers, students receive regular performance data in the form of informal classroom feedback, grades on assignments, examination scores, and end-of-course grades. The advent of standards-based testing regimes in American public education has increased dramatically the amount of information available to students, teachers, parents, and the public, particularly about students' mathematics and reading skills. Performance information that students receive throughout their educational careers can play an important role in educational investment decisions in several ways—by making the student eligible for specific educational interventions or policies that in turn affect additional schooling, by affecting the student's self-judgments, or by influencing the perceptions of parents, teachers, or peers, who in turn influence students.

First, such information can make students eligible for specific educational interventions or policies. There are clearly many such possibilities—eligibility for extra support, grade promotion, or advanced classes can all be based on test scores. We provide two examples relevant to our analysis in Massachusetts: high school exit examinations and merit scholarships. Today, approximately 70 percent of the nation's children, including those in Massachusetts, must pass exit examinations in order to graduate from high school (McIntosh 2012). In earlier work, we and others have examined the effects of barely failing these exit examinations (Martorell 2005; Papay, Murnane, and Willett 2010, 2014; Reardon et al. 2010; Ou 2010). In this paper, we focus on labels that hold no official consequences for students. Massachusetts also bases eligibility for its state-sponsored Adams Scholarship to support post-secondary education on students' standardized test scores. To be eligible, students must earn *Advanced* in either mathematics or ELA, at least *Proficient* in the other subject, and be in the top 25 percent of all test takers in their district in terms of total score (Cohodes and Goodman 2014). In this case, a label might induce college-going through financial aid. We discuss this scholarship in more detail below.

A second pathway through which performance information can influence educational attainments is by directly affecting students' perception of their abilities. Starting with Brookover, Thomas, and Paterson (1964), sociologists and psychologists have marshaled substantial evidence that students' self-judgments about their potential for academic success can affect educational outcomes (for example, Crocker et al. 2003, Shen and Pedulla 2000). Students' perceptions of their abilities are not static and can change rapidly over the course of their school careers. A number of scholars have hypothesized that adolescents' beliefs about their academic abilities, and therefore their educational decisions, may be particularly susceptible to new information. Aronson and Steele (2005) write: "Although clearly not the most fragile thing in nature, competence is much more fragile—and malleable—than we tend to think" (p. 436). Recent work in neuroscience supports this perspective, as it suggests that adolescents are making important decisions about investing in further schooling during a period in which their cognitive development is not yet complete and levels of hormones that affect judgments are changing rapidly (Steinberg 2008, 2010).

These psychological studies largely rely on experimental manipulations in a laboratory setting. However, several recent studies provide evidence that students do update their expectations about how much education they will complete when they receive new information about their academic performances (Jacob and Linkow 2011, Stinebrickner and Stinebrickner 2012).[1] Recent work has focused on decisions students make after they are enrolled in college, such as decisions to drop out or to choose a major (Stinebrickner and Stinebrickner 2013, 2014; Zafar 2011, 2013). These studies all demonstrate that students' educational decisions depend on new information they receive about their academic performance. The extent to which this information affects decisions about college-going itself remains less clear.[2]

Third, information can affect students indirectly by influencing the perspectives (and behavior) of their peers, parents, or teachers. Parents and teachers may reward or encourage students who score well. For example, a teacher looking through the list of

[1] While the sociological literature distinguishes between "educational aspirations" and "educational expectations," Jacob and Linkow (2011) show that the responses to survey questions aimed at capturing the two concepts are extremely highly correlated and very difficult to separate empirically. For that reason, we do not distinguish between these concepts and use "expectations" or "plans" throughout the paper.

[2] Several recent studies have shown that students' college-going decisions are sensitive to interventions even quite late in their educational careers. For example, Carrell and Sacerdote (2013), Hoxby and Turner (2013), and Castleman and Page (2014) all illustrate, using randomized field experiments, that interventions quite late in the senior year or in the summer after high school graduation increase substantially the actual college enrollments of low-performing or low-income students.

students who pass the test may see certain students in a new light because of their successes. There is a long literature in education that teachers' expectations of student performance affect student outcomes (for example, Rosenthal and Jacobson 1968, Jussim and Harber 2005). In fact, President George W. Bush argued that the No Child Left Behind Act was necessary in part because it challenged the "soft bigotry of low expectations" for disadvantaged and minority students.

Importantly, these indirect effects can take many forms and can be either reinforcing or compensatory. In other words, parents may reinforce the positive effect of earning a better label by talking to the student about college, or they may compensate for the negative effect of earning a worse label by providing the student with additional tutoring or supports. Obviously, these responses would operate in different directions, and we cannot disentangle them here.

While we cannot identify the mechanism at play, understanding whether performance data influences students' decisions about investing in post-secondary education is important for two related reasons. First, the amount of information that students receive, particularly from official state standardized tests, has grown tremendously over the past decades, and students, teachers, and schools are under increased pressure to improve results on these tests. Thus, the performance information embedded in these tests is ubiquitous and, perhaps, particularly salient.

Second, college-going behavior is especially important to examine given that the college wage premium is substantial and much greater than it was four decades ago (Autor 2014). College access has become a centerpiece of educational policy discussions in part because the growth in college enrollment rates has slowed in recent years. For example, both the U.S. Department of Education and the Gates Foundation have focused substantial attention—and resources—on college readiness and access. As a result, understanding its determinants, particularly in the K–12 educational context, is especially timely and policy relevant.

C. Research Questions

In this paper, we focus on performance labels that do not carry official consequences for students. We examine how students respond to a specific piece of information about their performance—the label that they earn on the Massachusetts standardized mathematics examination. Using a regression-discontinuity design, we examine the impact of the labeling by comparing the educational plans and attainments of students who were assigned exogenously to different labels because they scored close to, but fell on different sides of, the state-mandated labeling cut-points. We focus on testing in mathematics and for urban, low-income students, based on past work that looked at the effects of barely passing the high-stakes high school exit examination (Papay, Murnane, and Willett 2010).[3] We confirmed that there are no effects of labeling on the ELA

[3] We found that, for students on the margin of passing, barely passing the mathematics exit examination increased the probability of graduation substantially, but there were no effects of barely passing the ELA examination. Similarly, we found that the effects of barely passing the mathematics examination were concentrated among urban, low-income students. In related work, we have examined the effect of failing the mathematics and ELA exit examinations simultaneously using a multidimensional regression-discontinuity model (Papay, Willett, and Murnane 2011; Papay, Murnane, and Willett 2014). The statistical power requirements of such an approach preclude our using it here.

examination or for students who were higher income or from suburban schools on the low-stakes cutoffs we examine here.[4]

To assess whether performance labels affect educational outcomes more for some students than for others, we make use of students' responses to survey questions about their educational plans that were asked just before the students took the state examinations. The research community has long known that student plans predict educational attainment even after controlling for educational performance and other background characteristics (Duncan, Featherman, and Duncan 1972; Sewell, Haller, and Ohlendorf 1970; Sewell, Haller, and Portes 1969). However, the development of educational expectations is a process that researchers do not understand well (Jacob and Linkow 2011, Zafar 2011). We examine whether the effects of labeling depend on students' initial post-secondary educational plans and whether the performance labels students receive cause them to update these plans.

Importantly, we have framed the discussion in terms of students' responses to receiving a beneficial performance label—a result of obtaining a score just above a cut-point. However, students could also respond negatively to receiving a negative performance label—a result of obtaining a score just below a cut-point. We cannot distinguish the effect of "encouragement" from that of "discouragement" with our regression-discontinuity strategy. However, we use students' test score histories to provide some descriptive insight into which groups appear to experience encouragement effects and which appear to experience discouragement effects.

To summarize, our specific research questions are:

1. Does the performance label that urban, low-income students receive on the Massachusetts state mathematics test affect their college enrollment decisions?
2. Does the performance label affect intermediate outcomes such as test scores and the probability of high school graduation?
3. Are the post-secondary plans and college enrollment decisions of students who did not plan initially to attend a four-year college especially sensitive to the information embedded in the performance labels?
4. Does prior test performance shed light on the relative importance of encouragement and discouragement effects?

III. Research Design

A. Data Sources

Our data come from Massachusetts, a state that has placed a high priority on educational reform. Since the Massachusetts Education Reform Act of 1993, which introduced standards-based reforms and state-based testing, Massachusetts has invested substantially in K–12 public education. Under these reforms, the state began administering the Massachusetts Comprehensive Assessment System (MCAS) mathematics and English language arts (ELA) examinations in 1998. For most students, performance on these tests carries no official consequences. However, starting with the class of 2003, the tenth

[4] Detailed results for this and all findings mentioned in the paper are available from the first author on request.

grade tests became high-stakes exit examinations that students must pass in order to graduate from high school.[5]

To address our research questions, we have integrated several data sets provided by the Massachusetts Department of Elementary and Secondary Education. The first comes from the state's longitudinal data system, which tracks students throughout their school careers (K–12) and includes unique student identifiers, MCAS test results, demographic characteristics, school and district identifiers, and responses to surveys that students complete just before taking the MCAS examinations. We have supplemented this data set with records from the National Student Clearinghouse that track students' post-secondary educational attainments.

We focus on examinations and performance labels that have no official, state-determined consequences for students; in other words, they are "low stakes" from the perspective of the student. In eighth grade, the examination is used to hold schools and districts accountable but not students or individual teachers. However, the tenth grade examination is a high-stakes exit examination. As a result, in tenth grade we focus on students whose scores fall well above the passing cutoff. Specifically, we examine the effects of labeling at three different cutoffs: At *Needs Improvement* versus *Warning* on the eighth grade test, at *Proficient* versus *Needs Improvement* on the eighth and tenth grade tests, and at *Advanced* versus *Proficient* on the eighth and tenth grade tests. The state does not treat students on either side of these cutoffs differently and this information is not provided to colleges.[6] Furthermore, state officials report that the performance labels were likely not major factors in district policies because of the structure of the accountability system itself, which focuses on a Composite Performance Index rather than individual labels. Nonetheless, it is possible that individual districts or schools may have policies that determine students' eligibility for special programs or support services based on their performance labels.

We pool data across several years, examining students who took the eighth or tenth grade mathematics examinations in the Spring of 2003 through 2007. These students are members of the graduating cohorts of 2005–11. For each year, we restrict our sample to students who took the MCAS examination for the first time in that grade, excluding any students who had repeated the grade and were taking the test for a second time.

[5] Starting with the class of 2010, the state also included science examinations as part of the exit examination requirement. This requirement did not apply to students in our sample.

[6] Importantly, for some students, the performance label they earn in Grade 10 makes them eligible for a state-sponsored Adams Scholarship to support post-secondary education and is therefore a cutoff that carries consequences for students. To be eligible, students must earn *Advanced* in either mathematics or ELA, at least *Proficient* in the other subject, and be in the top 25 percent of all test takers in their district in terms of total score. Cohodes and Goodman (2014) explain this selection process in detail. Thus, for some students, scoring *Advanced* instead of *Proficient* or scoring *Proficient* instead of *Needs Improvement* in mathematics could make them eligible for the scholarship. To avoid any potential confounding of the effects of scholarship receipt and performance labeling, we focus our attention in tenth grade on the sample of students for whom earning these labels in mathematics does not affect their scholarship eligibility. We do this in two ways for each cutoff. For the *Advanced/Proficient* cutoff, we first exclude from our sample any student who scored *Proficient* on the ELA examination. Second, we exclude only students who scored *Proficient* on the ELA test and who scored in the top 25 percent of their district. Both samples give us quite similar results, so we present our findings using the less restrictive second sample. For the *Proficient/Needs Improvement* cutoff, we follow analogous restrictions, excluding students who scored *Advanced* on the ELA test and in the top 25 percent of their district.

B. Measures

To address our research questions, we created several outcome variables. Our primary outcome (*$COLL_i$*) measures whether students attended college within one year after their cohort's intended high school graduation date.[7] We also examine effects on intermediate outcomes, including students' tenth-grade mathematics test scores (standardized within grade and year), and whether they graduated from high school on time with their cohort. We created an additional outcome by recoding responses about post-secondary educational plans that students provide immediately before they take the MCAS examinations. The survey question to which they responded reads as follows:[8]

> Which of the following best describes your *current plans* for what you will do *after you finish high school*?
>
> A. I plan to attend a four-year college.
> B. I plan to attend a community college, business school, or technical school.
> C. I plan to work full-time after graduating from high school.
> D. I plan to join the military after graduating from high school.
> E. I have other plans.
> F. I have no plans right now.

The state has asked this question of all tenth graders since the 2002–2003 school year and of all eighth graders starting in 2005–2006. Seventy-six percent of Massachusetts' low-income urban tenth grade students (and 85 percent of eighth graders) completed this survey. We focus on four-year college plans and code a dichotomous outcome (*$COLL\_PLAN_i$*) to indicate whether the student reported that he or she planned to attend a four-year college after high school or not.[9]

Our key predictors come from the state testing data set, which includes test information at four levels: as item-level responses, raw scores, scaled scores, and performance levels. The scaled scores range from 200–280 in increments of two points, with a different performance rating each 20 points, as follows:

(a) 200–218: Failing/Warning
(b) 220–238: Needs Improvement

[7] We focus on enrollments after the student (or their cohort) graduated from high school in order to avoid mistakenly counting students who take college courses during high school as college enrollees. We define on-time cohort graduation as occurring four years after the student took the eighth grade examination and two years after the student took the tenth grade examination.

[8] Although the state has made minor changes to the question over the years, all versions are quite similar to the wording from the 2005 administration presented here.

[9] Note that we focus on four-year college plans here, but we define our measure of college attendance to include any student who entered a two-year or a four-year college. We make this distinction for several reasons. First, given that nearly all students in the state plan to attend some college, defining a demographic subgroup based on this distinction would not be particularly meaningful. Furthermore, expressing four-year college-going plans is actually a stronger predictor of college attendance (at either a two-year or four-year college) than plans to attend any college. Third, and more importantly, some students who enter a two-year college eventually matriculate to a four-year college. Unfortunately, we cannot track the progress for most students through post-secondary education long enough to observe this pattern, so we instead count students as attending college if they enter either a two-year or four-year college initially.

(c) 240–258: Proficient
(d) 260–280: Advanced

Because the scaled scores have such a coarse scale, with multiple raw scores mapping on to a single scaled score, we use raw scores in our analyses.[10]

To implement our regression-discontinuity approach, we center students' raw scores by subtracting out the value of the corresponding minimum score associated with the relevant cut-point. On the recentered continuous predictor ($MATH_i$), which serves as the "forcing variable" in our regression-discontinuity analyses, a student with a score of zero had achieved the minimum score to earn the more positive label. We also code a dichotomous version of this same predictor ($ABOVE_i = 1\{MATH_i \geq 0\}$).[11] To address our fourth research question, we include lagged versions of this last predictor ($PAST\_ABOVE_i$) to indicate whether the student fell above or below the relevant cut score on the previous test they took (for example, Grade 8 for the Grade 10 analyses).

C. Sample

The extent to which we can examine each of our outcomes for specific cohorts depends on the timing of the initial test and outcome data collection. In other words, we must have five years of data after the eighth grade test to examine the effect of classification on college outcomes but only two years to examine the effects on college-going plans expressed in tenth grade. As seen in Table 1, each of our analyses uses a different number of years of data. Importantly, because survey responses of eighth graders about post-secondary plans are only available beginning with the 2005–2006 cohort, we use data from only two cohorts for the analyses that examine whether the effects of eighth grade performance labels depend on students' initial college-going plans (our third research question), and we cannot examine actual college-going outcomes for this group.

As described above, we focus our presentation and discussion on students who are eligible for federal free or reduced price lunch programs and who are enrolled in one of Massachusetts's 22 urban school districts.[12] This group contains 18 percent of Massachusetts' eighth grade students who take the exam (and 13 percent of tenth graders). In several of our analyses, we examine heterogeneity based on the post-secondary educational plans that students report before they take the examinations. Among urban, low-income students, 63 percent of those who completed the tenth grade survey and 57 percent of those who completed the eighth grade survey reported that they planned to attend a four-year college. Thus, approximately 5–8 percent of students in the state are urban, low-income students who do not plan to attend a four-year college.

[10] Although multiple raw scores map to the same scaled score, each raw score corresponds to only one scaled score in a given year. For example, in 2004, all students earning 23, 24, 25, or 26 points received a 220. The state derives its scaled scores by using a piecewise nonlinear transformation that leads to clumping of students near the scaled score thresholds. As we illustrate below, there are no such issues with the raw scores. For more information on MCAS scoring and scaling, see the MCAS Technical Reports (Massachusetts Department of Elementary and Secondary Education 2002, 2005).

[11] In this presentation, we use the word "Above" to indicate students who earned the more positive label and "Below" to indicate students who earned the less positive label.

[12] The state defines urban districts as those that participate in the state's Urban Superintendents Network.

Table 1
Description of Data Structure and Cohorts, by the Availability of Specific Outcomes

Test Cohort	College-Going Plans	College Attendance
Eighth grade		
2002–2003	2004–2005	2007–2008
2003–2004	2005–2006	2008–2009
2004–2005	2006–2007	—
2005–2006	2007–2008	—
2006–2007	2008–2009	—
Tenth grade		
2002–2003	—	2005–2006
2003–2004	—	2006–2007
2004–2005	—	2007–2008
2005–2006	—	2008–2009

In Table 2, we describe our sample in more detail. In the first two columns, we present the number of urban, low-income students in each test-performance category along with the number who responded to the survey question concerning post-secondary educational plans. Note that these figures do not correspond to response rates for eighth graders because the survey questions were not asked every year. In the third column, we present the sample proportion of urban, low-income students who reported planning to attend a four-year college by performance level. Clearly, there is a strong performance gradient in college-going plans: Students with better scores have a much greater probability of planning to attend a four-year college than their lower-performing peers. For example, 85 percent of low-income urban tenth graders who score *Advanced* plan to attend a four-year college compared to just 46 percent of those who score *Failing*. However, that nearly half of students in the *Failing* category plan to attend a four-year college suggests that it is a popular option even for low-performing students.

Importantly, planning to attend a four-year college is not simply a catch all for a specific demographic group. In fact, among urban, low-income students, white males are the least likely to express plans to attend a four-year college. In Figure 1, we display the sample probability that low-income urban tenth grade students express plans to attend a four-year college by race and gender. Asian students have the highest probability of expressing four-year college plans followed closely by African-American students. Hispanic students express slightly higher probabilities than whites. Across all racial/ethnic groups, the sample probability that a girl plans to attend college is greater than for boys. Importantly, we find very similar patterns when we condition on student test scores.

D. Data Analyses

To examine the causal effect of performance labeling on post-secondary educational attainments, we use a regression-discontinuity strategy. By examining students immediately

Table 2

Number of Urban, Low-Income Students in the Sample, Number Reporting Post-Secondary Plans on Surveys Given before the Test, and the Percentage of Urban, Low-Income Students who Reported Plans to Attend a Four-Year College, by Performance Level on the Mathematics MCAS Test in Eighth Grade (2002–2003 to 2007–2008) and Tenth Grade (2002–2003 to 2006–2007)

Performance Label	Number of Urban, Low-Income Students	Number of Urban, Low-Income Students Responding to the Survey	Percent of Urban, Low-Income Students Planning to Attend a Four-Year College
Eighth grade			
Advanced	2,015	876	84.7
Proficient	8,068	3,358	74.4
Needs improvement	19,744	7,328	62.8
Warning	39,402	11,895	46.7
Tenth grade			
Advanced	4,699	4,014	84.9
Proficient	7,073	5,721	71.0
Needs improvement	11,854	9,337	60.0
Failing	12,573	8,486	45.6

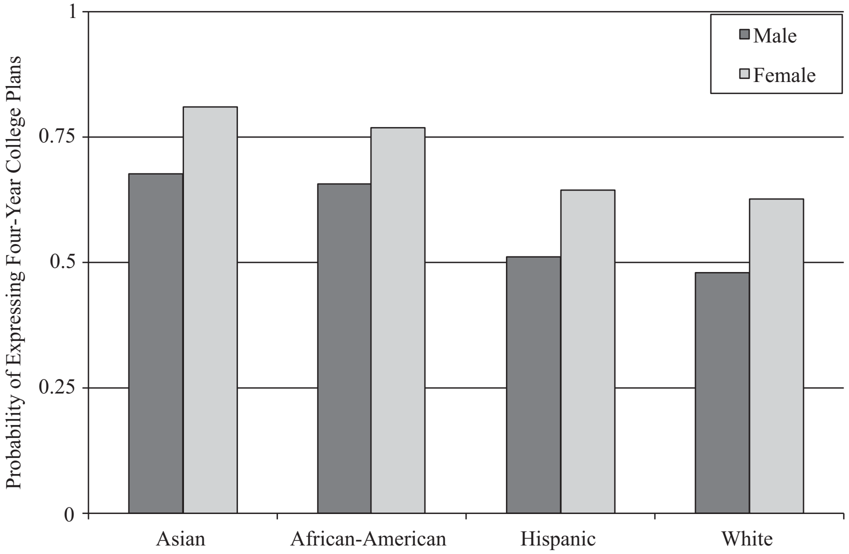


Figure 1

Sample Probabilities of Expressing an Interest in Attending a Four-Year College by Race and Gender for Urban, Low-Income Students in Massachusetts.

on either side of each cut score on the forcing variable, we compare the population probability of attending college for two groups of students—those who scored at the cut score and earned the more positive label (represented by parameter γ_{above}) and those (hypothetical) students who scored at the cut score yet received the less positive label (represented by parameter γ_{below}), as follows:

$$\gamma_{above} = \lim_{MATH_i \to 0^+} [P(COLL_i = 1) \mid MATH_i] \text{ and}$$

$$\gamma_{below} = \lim_{MATH_i \to 0^-} [P(COLL_i = 1) \mid MATH_i]$$

If the cut score was established exogenously, then students just on either side of the cut score must be equal in expectation prior to labeling; the estimated difference between these parameters provides an unbiased estimate of the causal impact of the classification for students in the population at the cut score (Lee and Lemieux 2010, Murnane and Willett 2011). Because the labels are applied rigidly such that all students who score below the cutoff on the forcing variable are assigned one label and all students who score above the cutoff are assigned a different—and more positive—label, our discontinuity is sharp.

In this presentation of the analytic method, we focus on the college-attendance outcome, but we use the same analytic strategy to conduct analyses of our other outcomes. In its basic formulation, this approach involves fitting a linear probability model of the following form:

$$(1) \quad p(COLL_i = 1) = \beta_0 + \beta_1 MATH_i + \beta_2 ABOVE_i + \beta_3 (ABOVE_i \times MATH_i) + \varepsilon_i$$

for the *i*th student.[13] In this model, β_2 represents the causal effect of interest. If its estimated value is statistically significant and positive, we can conclude that classifying a student at the cut score as earning the more positive label, as opposed to the less positive label, causes the student's probability of attending college to *increase* discontinuously, on average, in the population.

The internal validity of our regression-discontinuity analyses—and consequently our ability to make the required unbiased causal inferences—rests on several assumptions. The cut score must be determined exogenously, and students must not be able to manipulate knowingly their position on the forcing variable relative to it. Given the complicated scaling procedures used to determine the cutoffs, we have strong reason to believe that these assumptions hold, and we later present evidence that they do. We must also assume that we can model credibly the underlying relationship between the probability of attending college and the forcing variable, student MCAS score. Because our parameters of interest—γ_{above} and γ_{below}—represent limits projected onto the discontinuity from left and right, we estimate them using the nonparametric smoothing method of local linear regression implemented within an explicitly defined bandwidth on either side of the discontinuity, as recommended by Hahn, Todd, and Van der Klaauw (2001).[14]

[13] We used a linear probability, rather than a logistic or probit, specification of the hypothesized relationship between our outcome—a dichotomous indicator of college attendance—and predictors. As noted by Angrist and Pischke (2008), in large samples, the linear probability specification provides unbiased (consistent) estimates of the underlying trends while simplifying interpretation enormously. In addition, we use *local* linear-regression analysis—based only on observations within a narrow bandwidth—and so the linear specification of our statistical models is even more credible.

[14] Fan (1992) shows that, unlike most nonparametric smoothing techniques, local linear regression does not require boundary modifications.

To determine the amount of smoothing imposed during the local linear-regression analysis, we estimate an optimal value for the bandwidth (h^*) using a well-defined statistical fit criterion and a cross-validation procedure described by Imbens and Lemieux (2008).[15] We estimate h^* separately for each analysis and report these optimal bandwidths in our tables. To produce our figures, we fit local linear-regression trends using this bandwidth across the entire range of our data. However, our causal inferences derive only from estimates of projections onto the outcome axis at the cut score. As a result, we can estimate the parameter of principal interest in our analyses, the difference between γ_{above} and γ_{below}, in one step by fitting the single local linear-regression model presented in Equation 1 using observations that fall only within one bandwidth (h^*) on either side of the relevant cut score.[16]

In our analyses, we extend this basic analytic approach in several ways. First, to improve the precision of our estimation, we add to the model in Equation 1 a vector of covariates describing selected aspects of the student's background.[17] We also include the fixed effect of cohort to account for average differences in our outcome across years.[18] We present our main results from models that control for these background characteristics. However, we find it reassuring that the results from uncontrolled models are quite similar.

To address our third research question, we fit a statistical model similar to that specified in Equation 1 using a similar local linear-regression approach. In this case, though, we include the student's pretreatment self-reported college plans (*COLL_PLAN$_i$*) as a covariate and interact it with the main predictors, as follows:

$$\begin{aligned} p(COLL_i = 1) = {} & \beta_0 + \beta_1 MATH_i + \beta_2 ABOVE_i + \beta_3 (ABOVE_i \times MATH_i) \\ & + \beta_4 (MATH_i \times COLL\_PLAN_i) + \beta_5 (ABOVE_i \times COLL\_PLAN_i) \\ & + \beta_6 (ABOVE_i \times MATH_i \times COLL\_PLAN_i) + \beta_7 COLL\_PLAN_i + \varepsilon_i \end{aligned} \tag{2}$$

for the ith individual. Again, we obtain estimates of causal effects by fitting the model to observations that fall within one bandwidth on either side of the relevant cut score on the forcing variable. In this model, parameter β_2 represents the causal effect of receiving the more positive performance label on the population probability of attending college for students at the margin who *did not* plan to attend a four-year college. The linear combination of parameters, $\beta_2 + \beta_5$, represents the same effect for students who *did* plan

[15] $h^* = \arg\min_{h} \frac{1}{N} \sum_{i=1}^{N} (\hat{COLL}_i(h) - COLL_i)^2$, where $\hat{COLL}_i(h)$ is the predicted value using a bandwidth of h. In some cases, this function does not reach a clear global minimum over the range of plausible bandwidths; in these cases, we use the local minimum that produces the smallest bandwidth, sacrificing statistical power in an effort to reduce bias.

[16] In all cases, we adjust our standard errors to account for the discrete nature of our assignment variable by clustering observations, as recommended by Lee and Card (2008). We cluster observations at each score point.

[17] We include dichotomous predictors that describe student race, gender, limited English proficiency status, and special education status. We also include an indicator for whether the student was new to the state's public school system in the given year.

[18] We tested whether adding school fixed effects would increase the explanatory power of our models and found that it did not. We also found that the critical estimated parameters were not sensitive to the decision of whether to include school fixed effects in the corresponding model.

to attend a four-year college. For these analyses, we necessarily restrict our sample to low-income urban students who completed the survey. We also follow this same analytic approach to address our fourth research question by including an indicator of the student's past test performance (*$PAST\_ABOVE_i$*) as a covariate in the model and interacting it with predictors of interest.

Importantly, one key limitation of all analyses using a regression-discontinuity approach is that the results pertain only to students with scores close to the cut-points on the forcing variable. However, one strength of our study is that we can look for labeling effects at different cut-points and for students in both eighth and tenth grade. We find similar patterns at each grade level, although we do see evidence that some labels matter more than others and that different margins may be at play at different points in the test score distribution.

IV. Findings

A. Does the Performance Label Affect Post-Secondary Enrollment Decisions?

We find that earning a more positive performance label causes urban, low-income students to attend college at greater rates, at least at certain performance levels. The effects are small but important substantively. In Panel 1 of Table 3, we present the estimated causal effects of earning a more positive performance label on college enrollment at each of the cut scores. Being classified as *Needs Improvement* as opposed to *Warning* in eighth grade increases the fitted probability of enrolling in college by 2.1 percentage points ($p = 0.056$). Because only 38 percent of urban, low-income eighth graders scoring near the cutoff enroll in college within one year of cohort graduation, a 2.1 percentage point difference represents a substantial effect.

We find no effect of earning *Proficient* instead of *Needs Improvement* on college-going in either grade. Interestingly, this is the cutoff that is used to define Adequate Yearly Progress under No Child Left Behind. Teachers may pay special attention to increasing the test score performance of these "bubble kids" thereby dampening any test labeling effects (Booher-Jennings 2005, Neal and Schanzenbach 2010). Alternately, more moderate labels like *Proficient* and *Needs Improvement* may not be as meaningful to students, parents, or teachers; this is consistent with results presented by Zafar (2011) on students' GPAs.

We find that receiving the *Advanced* rather than the *Proficient* label on the tenth grade mathematics test increases the probability that urban, low-income students enroll in college by 5.1 percentage points ($p = 0.024$). Again, this is a large impact, considering that fewer than 60 percent of the urban, low-income students scoring near the cutoff enroll in college. In contrast, receiving *Advanced* rather than the *Proficient* label on the eighth grade mathematics test has no impact on the probability of college enrollment. One possible explanation for the difference between the eighth and tenth grade results concerns the test itself. Only 3.6 percent of Massachusetts' low-income, urban students earn an *Advanced* rating in eighth grade compared to 13 percent of tenth graders. Consequently, students scoring near the *Advanced/Proficient* cut-score in eighth grade are quite high performing.

Table 3

Estimated Effect of Earning the More Positive Performance Label at Different Cutoffs and on Different Outcomes for Urban, Low-Income Students Scoring near the Cut Point

Grade of Test	Needs Improvement/ Warning	Proficient/ Needs Improvement	Advanced/ Proficient
	Panel I: College Attendance		
Eighth grade	0.021~	0.001	0.007
	(0.009)	(0.028)	(0.035)
	$h=3$	$h=8$	$h=4$
	5,801	6,313	1,248
Tenth grade	N/A	0.008	0.051*
		(0.010)	(0.020)
		$h=6$	$h=8$
		8,280	4,171
	Panel II: High School Graduation (in four years)		
Eighth grade	0.028***	0.019	–0.006
	(0.004)	(0.014)	(0.023)
	$h=5$	$h=6$	$h=10$
	13,832	7,517	4,962
Tenth grade	N/A	0.021	–0.002
		(0.014)	(0.012)
		$h=3$	$h=9$
		6,035	7,103
	Panel III: Tenth Grade Mathematics Scores		
Eighth grade	0.032**	–0.006	0.022~
	(0.003)	(0.019)	(0.010)
	$h=2$	$h=6$	$h=6$
	7,974	11,695	5,155

Notes: ~, $p<0.10$; *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. Cell entries include the parameter estimate, standard error (in parentheses), optimal bandwidth used, sample size, and approximate p-value. All inferences from two-tailed hypothesis tests. Estimated effects from a local linear regression-discontinuity model from Equation 1 using observations within one bandwidth on either side of the cutoff, with the following control predictors: student race, gender, whether the student was new to the state, was currently or formerly classified as limited English proficient or required special education, and the fixed effect of cohort.

B. Does the Performance Label Affect Intermediate Outcomes such as Test Scores and High School Graduation?

In the second and third panels of Table 3, we present results of the effects of labeling on two intermediate outcomes: tenth grade mathematics test scores and high school graduation. These reflect potential mediators indicating effects of the label on investments during high school. Particularly at the low end of the test score distribution, we find substantial effects on these intermediate outcomes. For example, in eighth grade, receiving a label of *Needs Improvement* instead of *Warning* increases students' tenth grade test scores by 0.032 standard deviations ($p = 0.001$) and their probability of graduating from high school by 2.8 percentage points ($p < 0.001$). Again, we see no evidence of effects of labels at the *Proficient/Needs Improvement* cutoff. Also, while earning an *Advanced* label (instead of *Proficient*) in eighth grade increases tenth grade test scores by 0.022 standard deviations ($p = 0.054$), we see no effect of earning an *Advanced* label on high school graduation. This is potentially because students on the margin of earning this label are quite likely to graduate from high school. For example, more than 85 percent of low-income students scoring within two points of this cutoff in either eighth or tenth grade graduate from high school on time.

C. Are the Plans and Decisions of Students Without College-Going Plans Especially Sensitive to Performance Labels?

Our hypothesis that at least some of the response to performance labeling operates through students' perceptions of their own ability led us to consider heterogeneity within the urban, low-income group. Indeed, we find that performance labels matter much more for students who reported before they took the examination that they did not plan to attend a four-year college than for those with such plans.

Importantly, our analysis of the eighth grade test is limited by data availability. The state first administered the eighth grade survey to students in 2006, so we cannot examine college attendance directly. Instead, we must rely on a proxy: students' expressed educational plans in tenth grade. The results are striking for students who do not plan to attend a four-year college. As seen in Table 4, being classified as *Needs Improvement* instead of *Warning/Failing* on the eighth grade mathematics examination raises students' fitted probability of expressing four-year college as their intended post-secondary plan on the tenth grade survey by four percentage points ($p = 0.088$). Earning a *Proficient* label instead of *Needs Improvement* raises the probability of expressing four-year college-going plans by 6.2 percentage points, although this does not rise to even marginal levels of statistical significance ($p = 0.238$). Finally, scoring *Advanced* instead of *Proficient* in eighth grade increases the probability of expressing four-year college-going plans by 14 percentage points ($p = 0.074$). These effects are substantial for the group of high-performing eighth graders who do not plan to attend college. In all cases, there are no effects for students who reported before taking the examination that they plan to attend a four-year college.

These effects are both large and important given the strong relationship between students' college-going plans and their actual probability of enrolling in college. In fact, urban, low-income students who express plans to attend a four-year college on the tenth grade survey have estimated odds of enrolling in college that are nearly 3.5 times greater than the odds for similar students with lower educational expectations. These odds are still 2.4 times greater than the odds for students without college-going plans after controlling for students' test scores and other demographic characteristics. Thus, expectations are strong predictors of students' actual educational attainments.

We find very similar patterns on the tenth grade test. Again, performance labeling appears to be important particularly for students who did not plan to attend a four-year college. In particular, for these students being classified as *Advanced* rather than

Table 4
Estimated Effect of Earning the More Positive Performance Label on College-Going Behavior at Different Cutoffs for Urban, Low-Income Students Scoring Near the Cut Point, by Whether They Express Plans to Attend a Four-Year College After High School

Outcome	Students with College Plans	Students Without College Plans	Sample Size
Panel I: Eighth grade			
Needs improvement/warning cutoff:			
Express four-year college plans	0.009	0.040~	3,824
(grade 10)	(0.031)	(0.021)	$h=5$
Proficient/needs improvement cutoff:			
Express four-year college plans	–0.028	0.062	4,487
(grade 10)	(0.012)	(0.051)	$h=8$
Advanced/proficient cutoff:			
Express four-year college plans	0.007	0.137~	2,294
(grade 10)	(0.023)	(0.071)	$h=8$
Panel II: Tenth grade			
Proficient/needs improvement cutoff:			
College attendance	–0.002	0.013	6,609
	(0.014)	(0.032)	$h=6$
Advanced/proficient cutoff:			
College attendance	0.027	0.099**	3,316
	(0.035)	(0.033)	$h=8$

Notes: ~, $p<0.10$; *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. Cell entries include parameter estimates, standard errors (in parentheses), and approximate p-value. All inferences from two-tailed hypothesis tests. Estimated effects from a local linear regression-discontinuity model from Equation 2 using observations within one bandwidth on either side of the cutoff, with the following control predictors: student race, gender, whether the student was new to the state, was currently or formerly classified as limited English proficient or required special education, and the fixed effect of cohort.

Proficient increases the fitted probability that they will attend college by 9.9 percentage points ($p=0.010$). For these tenth graders on the margin, the estimated probability of attending college increases from approximately 39 percent to 49 percent simply by being labeled *Advanced* instead of *Proficient*.

In Figure 2, we present these results visually. In each panel, we include the sample probabilities of attending college (for tenth grade students, Panels A and B) or expressing four-year college-going plans (for eighth grade students, Panels C–E) for students with and without college-going plans before they took the respective test. We overlay the fitted values from our local linear-regression analysis.

These figures illustrate several important patterns. First, students who report as tenth graders that they plan to attend a four-year college do indeed attend college at a substantially greater rate than students with other plans (Panels A and B). Similarly, students who report as eighth graders that they plan to attend a four-year college have a much greater probability of reporting the same plans in tenth grade (Panels C–E). Second, many students do update their educational expectations over time. Many students who plan to attend college (top line) do not enroll and, more surprisingly, some students who do not plan to attend college (bottom line) do end up enrolling.

Third, across all three cut-scores, the label does not result in a clear disruption in the smoothed relationship for students with four-year college plans. In other words, the new information that students receive when they earn the more positive label does not seem to affect their probability of attending college. However, for students without four-year college plans, earning the more positive performance label increases substantially the probability of attending college or of expressing four-year college-going plans. This effect is seen in the sharp disruptions at the cut score. The effect of being classified as *Advanced* appears to be particularly large both in eighth and tenth grades. By contrast,

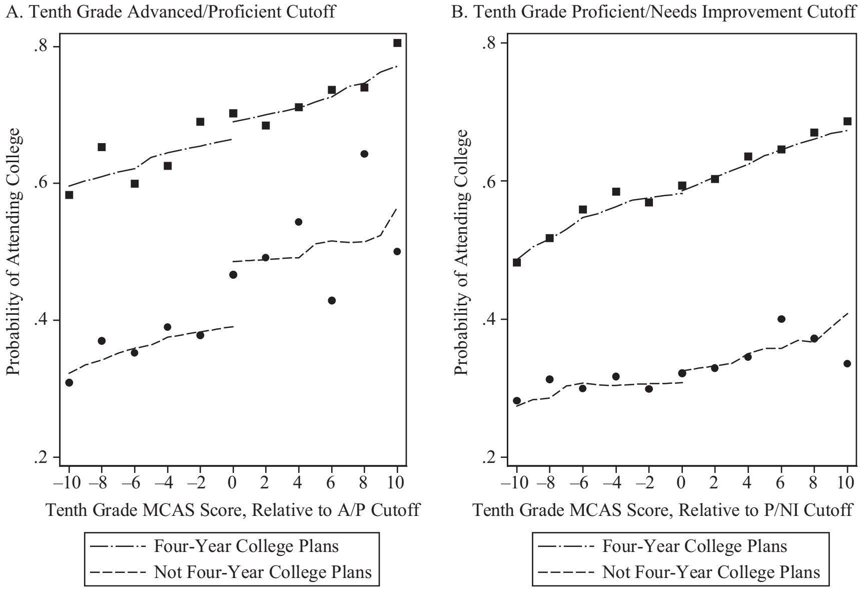
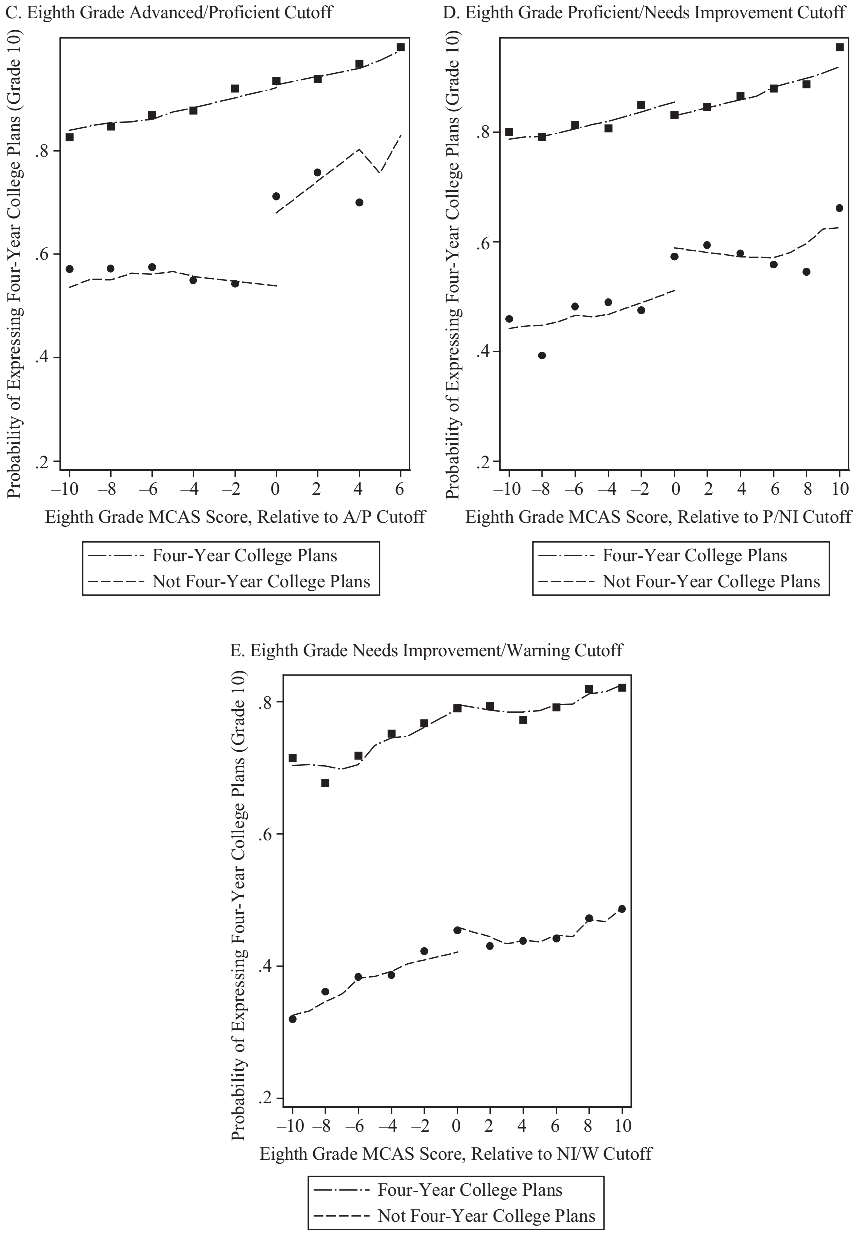


Figure 2

Fitted Local Linear-Regression Relationships between the Probability of Attending College and tenth Grade Mathematics Score Relative to the Advanced/Proficient Cutoff (Panel A, h = 8) and the Proficient/Needs Improvement Cutoff (Panel B, h* = 8),and the Probability of Expressing Four-Year College-Going Plans in Grade 10 and eighth Grade Mathematics Score Relative to the Advanced/Proficient Cutoff (Panel C, h* = 8), the Proficient/Needs Improvement Cutoff (Panel D, h* = 8), and the Needs Improvement/Warning Cutoff (Panel E, h* = 5), with the Sample Mean Probabilities Overlaid, for Urban, Low-Income Students who Do and Do Not Express Plans to Attend a Four-Year College before They Take the Test.*



C. Eighth Grade Advanced/Proficient Cutoff

D. Eighth Grade Proficient/Needs Improvement Cutoff

E. Eighth Grade Needs Improvement/Warning Cutoff

Figure 2 (*continued)*

while the effect of scoring *Proficient* instead of *Needs Improvement* appears to be greater for students without college plans than those with college-going plans, these differences are not statistically significant and the magnitudes are relatively small. Thus, labeling near the middle of the distribution appears to have less of an effect on students' postsecondary decisions than labeling at the top or bottom.

D. Does Prior Test Performance Shed Light on the Relative Importance of Encouragement and Discouragement Effects?

The findings presented so far indicate that the information embedded in performance labels affects students' college-going decisions. Students without plans to attend a four-year college are most liable to alter their decisions. However, from this analysis we cannot disentangle whether they reflect the positive effects of earning a better label or the negative effects of earning a worse label. For example, students who are labeled as *Advanced* could be encouraged by their performance, which could lead them to think more highly about their abilities, thereby increasing the probability they subsequently attend college. On the other hand, relatively high-performing students who are labeled as merely *Proficient* may be discouraged by their failure to achieve the more prestigious *Advanced* label and, as a result, may not consider themselves "college material." Unfortunately, since each group provides our estimate of the counterfactual for the other, our regression-discontinuity estimates cannot resolve this conundrum and only summarize the net effect.

In an attempt to shed light on the relative importance of encouragement and discouragement effects, we capitalize on information about students' past test performances. Unlike our regression-discontinuity analyses, which support strong causal inferences, this analysis is purely descriptive and is intended as a suggestive complement to the causal work presented above. Here, we assume that students, teachers, and parents respond to the information embedded in the test performance label when it is different from the label that students had earned in a previous grade. However, we assume that the responses are more limited if the new information matches students' prior labels. We present results from these analyses for the eighth grade test in Table 5, where we show the estimated causal effect of earning a more positive label for students who earned lower scores on their most recent test compared to those with higher scores on their most recent test. Suggestively, we find different patterns of responses at different parts of the test score distribution. For example, being labeled *Needs Improvement* instead of *Warning* has no effect for students who had scored *Warning* previously—this suggests that there is no encouragement effect for students near the bottom of the eighth grade test score distribution. However, we find substantial effects for students who had received a label of *Needs Improvement* or better in the past, which we interpret as a discouragement effect of earning *Warning* instead of *Needs Improvement*. At the middle and the top of the current test score distribution, the patterns are reversed, suggesting that earning a positive label encourages higher- performing students. On the tenth grade test, the performance level cutoffs are lower than on the eighth grade examination. For example, nearly all students near the *Advanced/Proficient* cutoff had scored *Proficient* or lower on the eighth grade test. Thus, in both eighth and tenth grades the encouragement effect appears to predominate at the top of the distribution.

Table 5

Estimated Effect of Earning the More Positive Performance Label on College-Going Behavior at Different Cutoffs, for Urban, Low-Income Students Scoring near the Cut Point, by whether they Scored Above or Below the Cutoff on an Earlier Test

Outcome	Students with Lower Scores on Prior Test	Students with Higher Scores on Prior Test	Bandwidth
Eighth grade			
Needs improvement/warning cutoff:			
Attend college	–0.006	0.108***	$h = 3$
	(0.032)	(0.022)	
	1,531	1,077	
Proficient/needs improvement cutoff:			
Attend college	0.061	–0.013	$h = 8$
	(0.035)	(0.042)	
	1,840	1,150	
Advanced/proficient cutoff:			
Attend college	0.086***	0.091	$h = 4$
	(0.013)	(0.050)	
	451	182	

Notes: ~, $p < 0.10$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. Cell entries include parameter estimates, standard errors (in parentheses), and approximate p-value. All inferences from two-tailed hypothesis tests. Estimated effects from a local linear regression-discontinuity model from Equation 2 using observations within one bandwidth on either side of the cutoff, with the following control predictors: student race, gender, whether the student was new to the state, was currently or formerly classified as limited English proficient or required special education, and the fixed effect of cohort.

V. Threats to Validity

As we have noted, the internal validity of a regression-discontinuity design depends on several important assumptions. First, the "treatment"—here embodied in the particular performance label applied to students scoring immediately above or below the cutoff—must have been assigned exogenously by the placement of the students with respect to the cutoff on the forcing variable and applied rigidly to all students. Second, students must not be able to manipulate their position on the forcing variable relative to the cut score. If these conditions hold, then all student characteristics, both observed and unobserved, should be a smooth function of the forcing variable around the cut score.

In our research, these assumptions hold because the cut scores differ from year-to-year based on a complicated scaling formula and are determined *after* students take the test; thus, it would be essentially impossible for students at the margin of passing knowingly to manipulate their explicit positions with respect to the cutoff while taking the examination. However, we can also test if these assumptions are met in several ways. First, if students can influence their position on the forcing variable relative to the cutoff by manipulating their test performance, we would expect the test score distribution to be

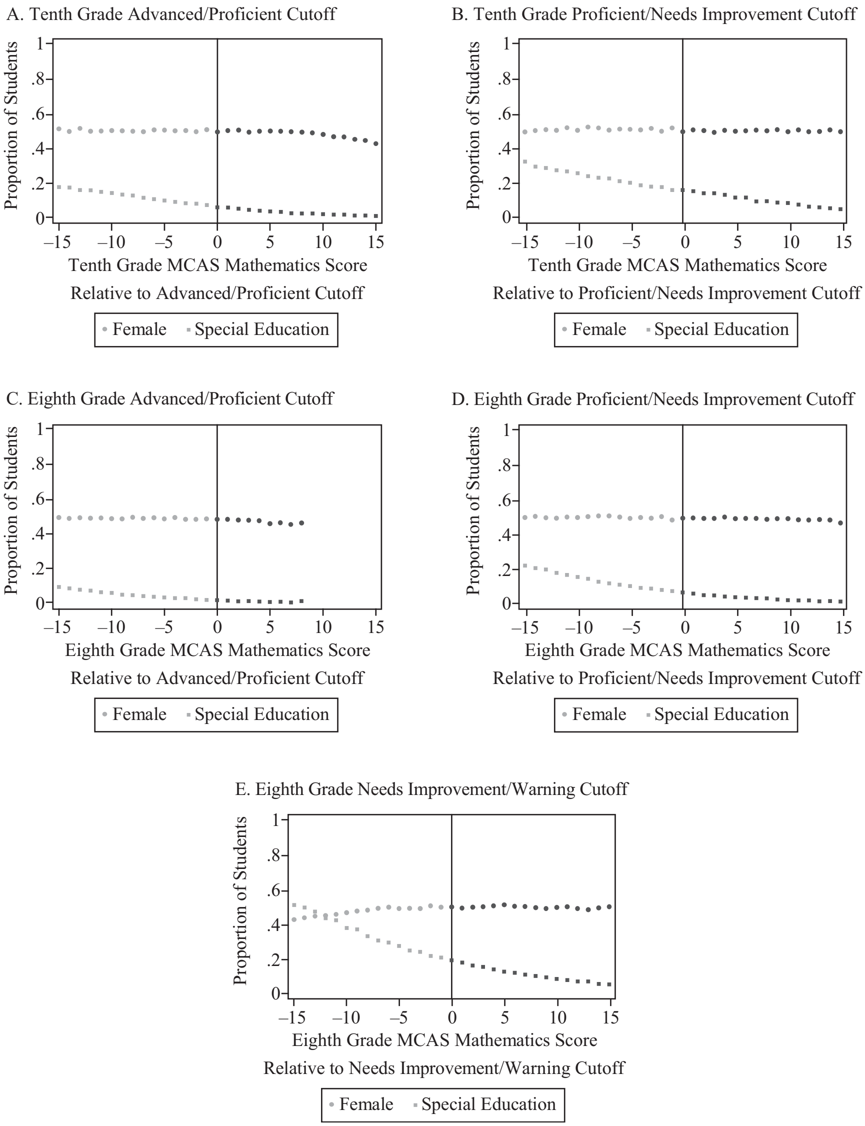


Figure 3

Proportion of Students who are Female and who Qualify for Special Educational Services by Test Score Relative to Each of the Cutoffs

Table 6

Results from the Hypothesis Test that the Disruption in each Observed Covariate is Zero in the Population at each of the Three Cut-Scores, from a Seemingly Unrelated Regression (SUR) Regression-Discontinuity Model where each Covariate is Treated as an Outcome, for the Full Sample of Urban, Low-Income Students and the Subsample of Students who Plan to Attend a Four-Year College

	Urban, Low-Income Students	Urban, Low-Income Students Without Four-Year College Plans
Panel I: Grade 8		
Needs Improvement/Warning cutoff	$\chi^2(11)=11.15$ $p=0.431$	$\chi^2(11)=6.63$ $p=0.829$
Proficient/Needs Improvement cutoff	$\chi^2(11)=10.28$ $p=0.505$	$\chi^2(11)=9.02$ $p=0.620$
Advanced/Proficient cutoff	$\chi^2(11)=6.55$ $p=0.835$	$\chi^2(11)=10.60$ $p=0.477$
Panel II: Grade 10		
Proficient/Needs Improvement cutoff	$\chi^2(11)=7.59$ $p=0.749$	$\chi^2(11)=6.90$ $p=0.807$
Advanced/Proficient cutoff	$\chi^2(11)=15.66$ $p=0.154$	$\chi^2(11)=9.96$ $p=0.534$

Notes: All inferences from two-tailed hypothesis tests. Covariates treated as outcomes include student race, gender, whether the student was new to the state, was currently or formerly classified as limited English proficient, or required special education.

discontinuous near the cutoff. In all cases, we see no discontinuity apparent at the cut score. Second, if students can influence their labeling after taking the test, we would expect to see some noncompliers, that is, students whose test scores fell below the cutoff but earned the more positive label regardless. Again, we see no cases of such manipulation in the data.

Third, we conducted exploratory analyses to check for smoothness in the relationship between observed student characteristics and the forcing variable around the cutoff. Visual inspection of the underlying distributions suggests that these relationships are smooth over the cut score. As an example, Figure 3 presents the average value of two covariates, the proportion of students who are female and who qualify for special educational services, based on student test scores.[19] In each case, we see that the relationship between the covariate and student test scores is smooth over the cutoff.

[19] We chose these two covariates because they were the two predictors that had the largest *t*-statistics in a model that regressed college going on our full set of covariates. We present these figures as an illustration, but we observe similar patterns among other covariates.

Table 7

Estimated Effect of Earning the More Positive Performance Label on College-Going Behavior at Different Cutoffs, by Bandwidth

	Bandwidth				
Outcome	h^*–2	h^*–1	h^*	h^*+1	h^*+2
Eighth grade *Needs Improvement/Warning* cutoff – all urban, low-income students:					
College attendance ($h^*=3$)	—	—	0.021 ∽	0.023 ∽	0.028 ∽
			(0.009)	(0.012)	(0.014)
Eighth grade *Needs Improvement/Warning* cutoff – urban, low-income students without four-year college plans:					
Express college plans (grade 10) ($h^*=5$)	–0.007	0.028	0.040 ∽	0.023	0.027
	(0.029)	(0.027)	(0.021)	(0.023)	(0.021)
Eighth grade *Proficient/Needs Improvement* cutoff – all urban, low-income students:					
College attendance ($h^*=7$)	–0.013	–0.001	0.004	0.001	0.003
	(0.030)	(0.030)	(0.030)	(0.028)	(0.027)
Eighth grade *Proficient/Needs Improvement* cutoff – urban, low-income students without four-year college plans:					
Express college plans (grade 10) ($h^*=8$)	0.109 ∽	0.074	0.062	0.086 ∽	0.085*
	(0.051)	(0.058)	(0.051)	(0.045)	(0.039)
Eighth grade *Advanced/Proficient* cutoff – all urban, low-income students:					
College attendance ($h^*=9$)	—	–0.047	0.007	0.000	0.012
		(0.024)	(0.035)	(0.026)	(0.027)

Eighth grade *Advanced/Proficient* cutoff – urban, low-income students without four-year college plans:					
Express college plans (grade 10) ($h^* = 8$)	0.148 ~	0.151 ~	0.137 ~	0.153*	0.123
	(0.076)	(0.072)	(0.071)	(0.069)	(0.073)
Tenth grade *Proficient/Needs Improvement* cutoff – all urban, low-income students:					
College attendance ($h^* = 6$)	0.031 ~	0.02 ~	0.008	–0.003	–0.003
	(0.014)	(0.010)	(0.010)	(0.012)	(0.011)
Tenth grade *Proficient/Needs Improvement* cutoff – urban, low-income students without four-year college plans:					
College attendance ($h^* = 6$)	0.048	0.024	0.013	0.002	0.016
	(0.043)	(0.035)	(0.032)	(0.032)	(0.030)
Tenth grade *Advanced/Proficient* cutoff – all urban, low-income students:					
College attendance ($h^* = 8$)	0.016	0.043 ~	0.051*	0.038 ~	0.030
	(0.021)	(0.022)	(0.020)	(0.019)	(0.018)
Tenth grade *Advanced/Proficient* cutoff – urban, low-income students without four-year college plans:					
College attendance ($h^* = 8$)	0.057	0.097*	0.099**	0.061	0.074*
	(0.040)	(0.037)	(0.033)	(0.036)	(0.030)

Notes: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. All inferences from two-tailed hypothesis tests. Estimated effects from a local linear regression-discontinuity model from Equation 1 using observations within one bandwidth on either side of the cutoff, with the following control predictors: student race, gender, whether the student was new to the state, was currently or formerly classified as limited English proficient or required special education, and the fixed effect of cohort.

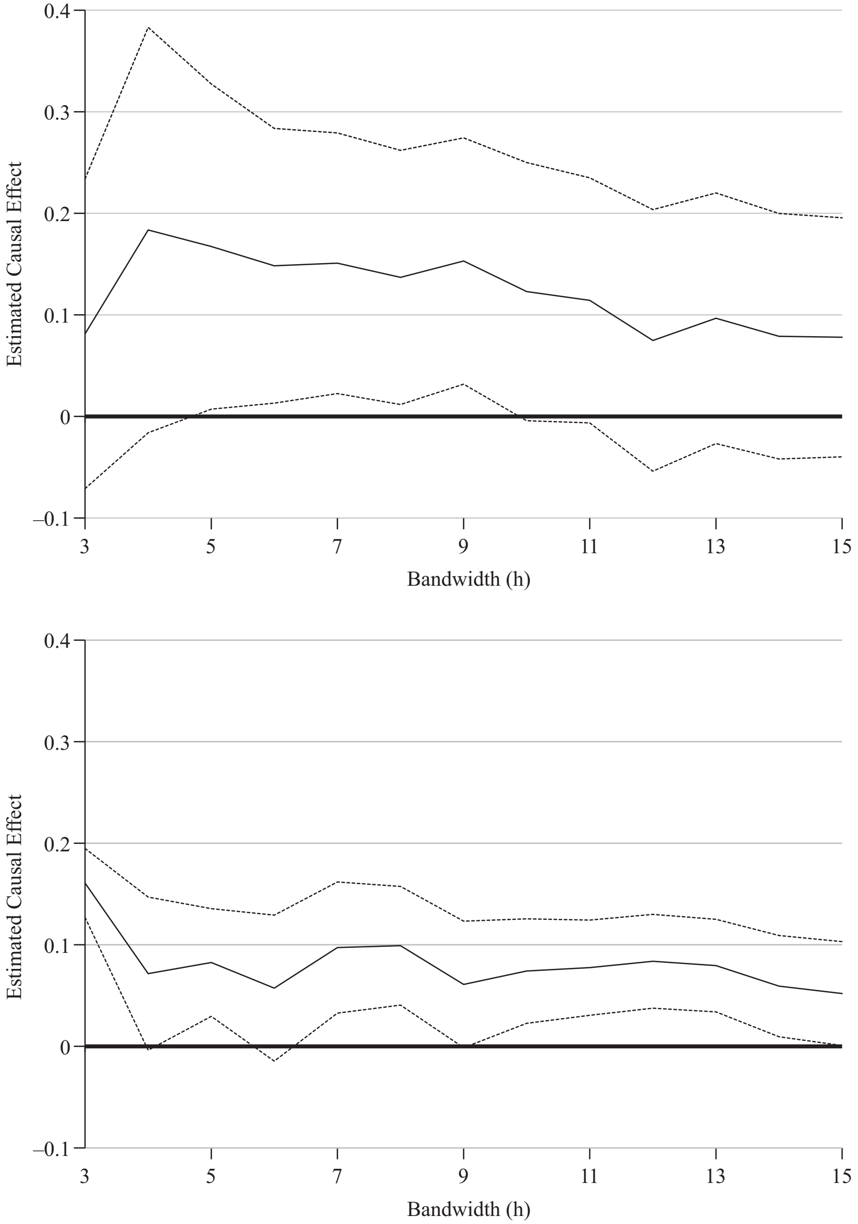


Figure 4

Estimated Causal Effect of Earning Advanced instead of Proficient on the eighth Grade (Top Panel) and tenth Grade (Bottom Panel) Tests on College-going from Local Linear Regression-Discontinuity Analysis with Bandwidths Ranging from three to 15 Score Points, with 90 Percent Confidence Intervals.

Seeking to summarize these analyses in a single test, we followed the approach suggested by Lee and Lemieux (2010). We fit a set of seemingly unrelated regression (SUR) models, each of which consists of our basic regression-discontinuity model but with a different covariate treated as the outcome. Then, we tested whether the coefficients on the discontinuity term equaled zero jointly across all covariates. We executed this procedure five times, once at each cutoff and grade. In Table 6, we present the results of these analyses both for our full sample of urban, low-income students and for the sample of students who did not plan to attend a four-year college. In all cases, we fail to reject the null hypothesis. This pattern supports our belief that the state has imposed the cut score exogenously.

The other key assumption underpinning our regression-discontinuity strategy is that we have specified the hypothesized relationship between the outcome and the forcing variable (mathematics test score) correctly, at least in the immediate vicinity of the cut score. We have addressed this issue by focusing our analyses within a narrow bandwidth around the cut scores on the forcing variables and adopting a flexible local linear-regression approach. The key decision in this analysis is the choice of bandwidth, *h*, which governs the amount of smoothing. To assess the sensitivity of our findings to this decision, we refitted our principal statistical models while restricting the sample to students whose test scores fell within different bandwidths around the cutoff on the forcing variable.

In Table 7, we present the estimated causal effects for each of our analyses across a range of bandwidths. Our main findings remain robust to the choice of bandwidth. While the magnitudes of a few individual estimates are sensitive to these choices, the general patterns persist across a range of bandwidths. In particular, the effects of earning *Advanced* instead of *Proficient* for students who do not plan to attend a four-year college are generally large, positive, and statistically significant. In Figure 4, we explore this result in more depth for both the eighth grade (top panel) and tenth grade (bottom panel) tests. Here, we present the estimated causal effect from our local linear-regression analysis with bandwidths ranging from 3–15 score points along with 90 percent confidence intervals. We see that, in both cases, the estimated effect of earning a more positive label is consistently large across a wide range of bandwidths. Given the smaller sample size in eighth grade, some of these estimates do not reach traditional levels of statistical significance, but the general pattern remains.

VI. Discussion

We conclude that performance labels on the state's mathematics test do indeed affect the college-going decisions of urban low-income students. Receiving a positive performance label, even on low-stakes tests that carry no official consequences for students, increases the probability that urban, low-income students attend college. The effect is particularly large for students who reported before they took the test that they did not plan to attend a four-year college.

There are three key features of this result that are particularly interesting. First, these labels matter even though they provide no additional information beyond the fine-grained test scores on which they are based. In other words, if all possible information were used, we should see no disruption in outcomes at the labeling cut scores. There are several complementary explanations for this phenomenon. First, some agent may be basing decisions and behaviors on the performance label and not the underlying test scores. For example, local policies may base eligibility for particular programs on performance labels. Or, students, parents, or teachers may rely on the performance labels as summaries instead of examining the finer-grained test scores. A second explanation is that the labels may evoke emotional responses. There is a growing literature in economics that focuses on the role of emotions and other psychological features in the decision-making process (for example, Kaufman 1999, Loewenstein 2001, Akerlof and Kranton 2002, Muramatsu and Hanoch 2005). Receiving performance labels like *Advanced* or *Warning* on a test that teachers and other adults have identified as important may well affect students, particularly adolescents whose cognitive processes are fragile and still in development.

A second key feature of our results is that effects of performance labels are concentrated among a particularly disadvantaged group of urban, low-income students. However, the relatively rare data on college plans permits us to explore a dimension of heterogeneity that is not part of standard student demographic profiles. We find that these effects are particularly large for students who do not plan to attend a four-year college after high school. The college enrollment decisions of this group appear to be especially sensitive to performance labels. We find no such effects of labeling for higher-income students or students attending suburban schools. Although only about 8 percent of students in the state are urban, low-income students without college-going plans, this group is a particularly important one for policymakers given the very high economic payoff to education credentials and the concerns nationally about upward mobility and educational inequality.

A third key feature is that labeling appears to matter more for students at the top or the bottom of the test score distribution. This suggests two complementary explanations. First, consistent with Zafar (2011), extreme performance information such as labels like *Warning* or *Advanced* may be especially salient to students or to their teachers. Alternately, schools may emphasize getting students over the proficiency cutoff and keeping them there (Neal and Schanzenbach 2010). To the extent that schools focus attention on “bubble kids” on either side of the *Proficient* cutoff, we would expect any effects of labeling at this middle cut-point to be muted.

Importantly, although we find that the label itself matters, we cannot determine precisely the mechanisms through which these effects operate. For example, students could respond directly, feeling encouraged or discouraged as a result of their performance. Parents, teachers, or school policies may also respond, producing indirect effects on student outcomes. Our results do suggest that at least part of the effect operates through student—rather than teacher—responses. The reason is that the effects of performance labels on students’ subsequent educational decisions are stronger for students without four-year college expectations even in models with school fixed effects.[20] Since teachers are unlikely to know directly about students’ college expectations, this pattern is inconsistent with the hypothesis that the observed effects stem primarily from teachers’ responses to students’ performance labels. At a minimum, any such effects must interact with the attitudes and behaviors of the students themselves.

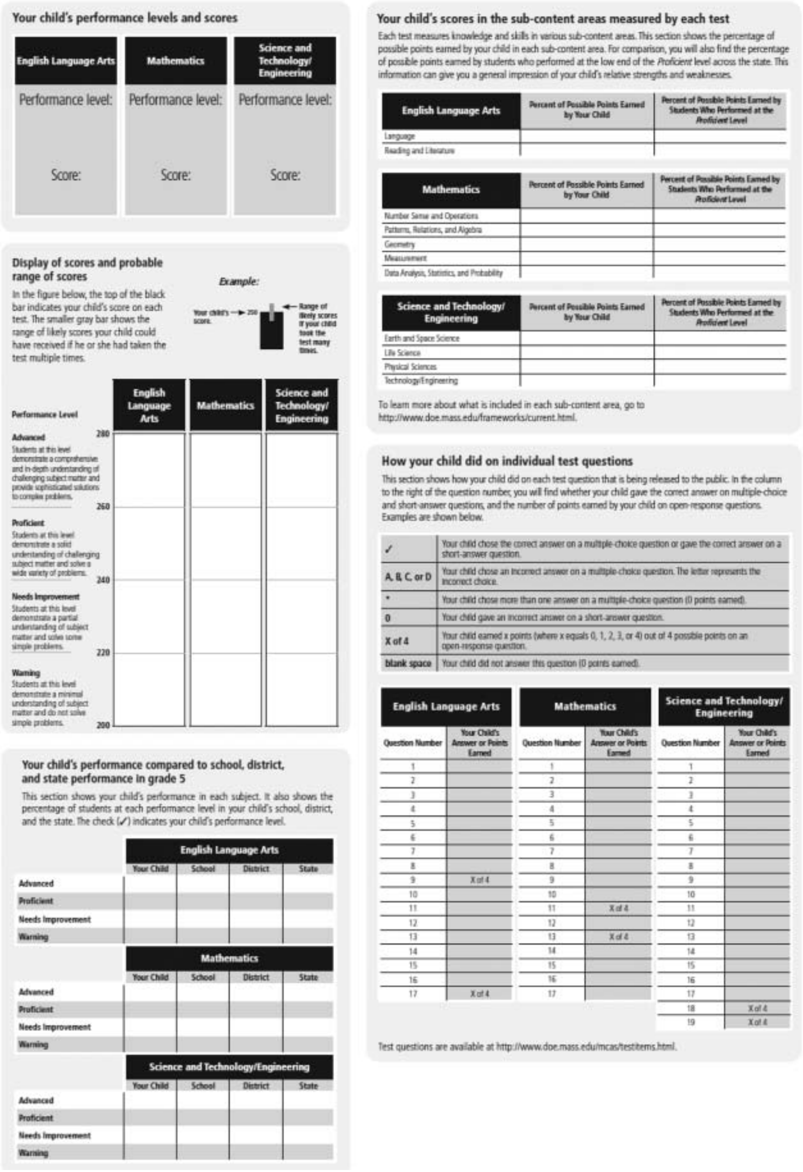
[20] Results from models with school fixed effects produce quite similar point estimates to those presented here, although the estimates are somewhat less precise.

In addition to the causal evidence about how students update their educational investment decisions, our findings have an important methodological implication for research that aims to identify the causal effects of policy interventions using a regression-discontinuity strategy. Often researchers take advantage of policies that assign students to treatment based on whether their value on a continuous forcing variable such as a test score falls below (or above) a particular cutoff. If individuals respond to performance labels on these tests, then estimates of the intervention's effects will be confounded with the effect of the labeling itself. In short, our paper presents evidence that mechanisms, including emotional responses, may be at play when students are assigned to groups based on test score performance. As a result, using such test score classifications as an exogenous source of assignment to treatments may produce biased estimates of the relevant treatment effects. In all cases, researchers must think carefully about the range of pathways through which assignment to treatment in a quasi-experimental design may affect student outcomes other than through the treatment itself.

Finally, this paper has substantive implications for policymakers. The labels we examine provide information about student performance relatively late in their academic careers. That students respond to these labels suggests that high school is not too late for educational interventions that influence college-going. The fact that dividing a continuous performance distribution into discrete categories affects the post-secondary educational enrollment decisions of urban, low-income students is clearly an important, unintended consequence of state testing policies as they have been implemented. Of particular concern is the finding that receiving a lower label appears to discourage students at the bottom of score distribution. Finding ways to support such students seems especially important. In order to formulate clear policy responses to the evidence that performance labels matter, though, it is important to learn whose behaviors are responding to the labels and the specific mechanisms through which the labels affect subsequent educational outcomes.

Appendix

Sample Report Provided to Students with their MCAS Test Results



Your child's performance levels and scores

English Language Arts	Mathematics	Science and Technology/ Engineering
Performance level:	Performance level:	Performance level:
Score:	Score:	Score:

Display of scores and probable range of scores

In the figure below, the top of the black bar indicates your child's score on each test. The smaller gray bar shows the range of likely scores your child could have received if he or she had taken the test multiple times.

Example:

Your child's score. → 250

← Range of likely scores if your child took the test many times.

Performance Level		English Language Arts	Mathematics	Science and Technology/ Engineering
	280			
Advanced Students at this level demonstrate a comprehensive and in-depth understanding of challenging subject matter and provide sophisticated solutions to complex problems.				
	260			
Proficient Students at this level demonstrate a solid understanding of challenging subject matter and solve a wide variety of problems.				
	240			
Needs Improvement Students at this level demonstrate a partial understanding of subject matter and solve some simple problems.				
	220			
Warning Students at this level demonstrate a minimal understanding of subject matter and do not solve simple problems.				
	200			

Your child's performance compared to school, district, and state performance in grade 5

This section shows your child's performance in each subject. It also shows the percentage of students at each performance level in your child's school, district, and the state. The check (✓) indicates your child's performance level.

English Language Arts	Your Child	School	District	State
Advanced				
Proficient				
Needs Improvement				
Warning				

Mathematics	Your Child	School	District	State
Advanced				
Proficient				
Needs Improvement				
Warning				

Science and Technology/Engineering	Your Child	School	District	State
Advanced				
Proficient				
Needs Improvement				
Warning				

Your child's scores in the sub-content areas measured by each test

Each test measures knowledge and skills in various sub-content areas. This section shows the percentage of possible points earned by your child in each sub-content area. For comparison, you will also find the percentage of possible points earned by students who performed at the low end of the *Proficient* level across the state. This information can give you a general impression of your child's relative strengths and weaknesses.

English Language Arts	Percent of Possible Points Earned by Your Child	Percent of Possible Points Earned by Students Who Performed at the *Proficient* Level
Language		
Reading and Literature		

Mathematics	Percent of Possible Points Earned by Your Child	Percent of Possible Points Earned by Students Who Performed at the *Proficient* Level
Number Sense and Operations		
Patterns, Relations, and Algebra		
Geometry		
Measurement		
Data Analysis, Statistics, and Probability		

Science and Technology/ Engineering	Percent of Possible Points Earned by Your Child	Percent of Possible Points Earned by Students Who Performed at the *Proficient* Level
Earth and Space Science		
Life Science		
Physical Sciences		
Technology/Engineering		

To learn more about what is included in each sub-content area, go to http://www.doe.mass.edu/frameworks/current.html.

How your child did on individual test questions

This section shows how your child did on each test question that is being released to the public. In the column to the right of the question number, you will find whether your child gave the correct answer on multiple-choice and short-answer questions, and the number of points earned by your child on open-response questions. Examples are shown below.

✓	Your child chose the correct answer on a multiple-choice question or gave the correct answer on a short-answer question.
A, B, C, or D	Your child chose an incorrect answer on a multiple-choice question. The letter represents the incorrect choice.
*	Your child chose more than one answer on a multiple-choice question (0 points earned).
0	Your child gave an incorrect answer on a short-answer question.
X of 4	Your child earned x points (where x equals 0, 1, 2, 3, or 4) out of 4 possible points on an open-response question.
blank space	Your child did not answer this question (0 points earned).

English Language Arts		Mathematics		Science and Technology/ Engineering	
Question Number	Your Child's Answer or Points Earned	Question Number	Your Child's Answer or Points Earned	Question Number	Your Child's Answer or Points Earned
1		1		1	
2		2		2	
3		3		3	
4		4		4	
5		5		5	
6		6		6	
7		7		7	
8		8		8	
9	X of 4	9		9	
10		10		10	
11		11	X of 4	11	
12		12		12	
13		13	X of 4	13	
14		14		14	
15		15		15	
16		16		16	
17	X of 4	17		17	
				18	X of 4
				19	X of 4

Test questions are available at http://www.doe.mass.edu/mcas/testitems.html.

Source: Massachusetts Department of Elementary and Secondary Education

References

Akerlof, George A., and Rachel E. Kranton. 2002. “Identity and Schooling: Some Lessons for the Economics of Education.” *Journal of Economic Literature* 40(4):1167–201.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton: Princeton University Press.

Aronson, Joshua, and Claude M. Steele. 2005. “Stereotypes and the Fragility of Academic Competence, Motivation, and Self-Concept.” In *Handbook of Competence and Motivation*, ed. Andrew J. Elliot and Carol S. Dweck, 436–56. New York: Guilford Press.

Autor, David. 2014. “Skills, Education, and the Rise of Earnings Inequality among the ‘Other 99 Percent’.” *Science* 344(6186):843–51.

Booher-Jennings, Jennifer. 2005. “Below the Bubble: ‘Educational Triage’ and the Texas Accountability System.” *American Educational Research Journal* 42(2):231–68.

Brookover, Wilbur B., Shailer Thomas, and Ann Paterson. 1964. “Self-Concept of Ability and School Achievement.” *Sociology of Education* 37:271–78.

Carrell, Scott E., and Bruce Sacerdote. 2013. “Late Interventions Matter Too: The Case of College Coaching New Hampshire.” NBER Working Paper 19031.

Castleman, Benjamin L., and Lindsay C. Page. 2014. *Summer Melt: Supporting Low-Income Students Through the Transition to College*. Cambridge: Harvard Education Press.

Cohodes, Sarah, and Joshua Goodman. 2014. “Merit Aid, College Quality, and College Completion: Massachusetts’ Adams Scholarship as an In-Kind Subsidy.” *American Economic Journal: Applied Economics* 6(4):251–85.

Crocker, Jennifer, Andrew Karpinski, Diane M. Quinn, and Sara K. Chase. 2003. “When Grades Determine Self-Worth: Consequences of Contingent Self-Worth for Male and Female Engineering and Psychology Majors.” *Journal of Personality and Social Psychology* 85(3):507–16.

Duncan, Otis D., David L. Featherman, and Beverly Duncan. 1972. *Socioeconomic Background and Achievement*. New York: Seminar Press.

Fan, Jianqing. 1992. “Design-Adaptive Nonparametric Regression.” *Journal of the American Statistical Association* 87(420):998–1004.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design.” *Econometrica* 69(1):201–209.

Hoxby, Caroline, and Sarah Turner. 2012. “Expanding College Opportunities for High-Achieving, Low Income Students.” Stanford Institute for Economic Policy Research Discussion Paper 12-014.

Imbens, Guido W., and Thomas Lemieux. 2008. “Regression Discontinuity Designs: A Guide to Practice.” *Journal of Econometrics* 142(2):615–35.

Jacob, Brian A., and Tamara W. Linkow. 2011. “Educational Expectations and Attainment.” In *Whither Opportunity? Rising Inequality, Schools, and Children’s Life Chances*, ed. Greg J. Duncan and Richard J. Murnane, 133–64. New York: Russell Sage Foundation.

Jussim, Lee, and Kent D. Harber. 2005. “Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies.” *Personality and Social Psychology Review* 9(2):131–55.

Kaufman, Bruce E. 1999. “Emotional Arousal as a Source of Bounded Rationality.” *Journal of Economic Behavior and Organization* 38(2):135–44.

Lee, David S., and David Card. 2008. “Regression Discontinuity Inference with Specification Error.” *Journal of Econometrics* 142(2):655–74.

Lee, David S., and Thomas Lemieux. 2010. “Regression Discontinuity Designs in Economics.” *Journal of Economic Literature* 48(2):281–355.

Loewenstein, George. 2001. “Preferences, Behavior, and Welfare: Emotions in Economic Theory and Economic Behavior.” *American Economic Review: AEA Papers and Proceedings* 90 (2):426–32.

Martorell, F. 2005. “Does Failing a High School Graduation Exam Matter?” Working paper.

Massachusetts Department of Education. 2002. “2001 MCAS Technical Report.”

Massachusetts Department of Education. 2005. “2004 MCAS Technical Report.”

McIntosh, Shelby. 2012. “State High School Exit Exams: A Policy in Transition.” Washington, D. C.: Center on Education Policy, George Washington University.

Muramatsu, Roberta, and Yaniv Hanoch. 2005. “Emotions as a Mechanism for Boundedly Rational Agents: The Fast and Frugal Way.” *Journal of Economic Psychology* 26(2): 201–21.

Murnane, Richard J. 2013. “U.S. High School Graduation Rates: Patterns and Explanations.” *Journal of Economic Literature* 51(2):370–422.

Murnane, Richard J., and John B. Willett. 2011. *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* New York: Oxford University Press.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. “Left Behind by Design: Proficiency Counts and Test-Based Accountability.” *Review of Economics and Statistics* 92(2):263–83.

Papay, John P., Richard J. Murnane, and John B. Willett. 2010. “The Consequences of High School Exit Examinations for Low-Performing Urban Students: Evidence from Massachusetts.” *Educational Evaluation and Policy Analysis* 32(1):5–23.

———. 2014. “High-School Exit Examinations and the Schooling Decisions of Teenagers: Evidence from Regression-Discontinuity Approaches.” *Journal of Research on Educational Effectiveness* 7(1):1–27.

Papay, John P., John B. Willett, and Richard J. Murnane. 2011. “Extending the Regression-Discontinuity Approach to Multiple Assignment Variables.” *Journal of Econometrics* 161 (2):203–207.

Rabin, Matthew, and Joel L. Schrag. 1999. “First Impressions Matter: A Model of Confirmatory Bias.” *Quarterly Journal of Economics* 114(1):37–82.

Reardon, Sean F., Nicole Arshan, Allison Atteberry, and Michal Kurlaender. 2010. “Effects of Failing a High School Exit Exam on Course Taking, Achievement, Persistence, and Graduation.” *Educational Evaluation and Policy Analysis* 32(4):498–520.

Sewell, William H., Archibald O. Haller, and George W. Ohlendorf. 1970. “The Educational and Early Occupational Status Attainment Process: Replication and Revision.” *American Sociological Review* 35(December):1014–27.

Sewell, William H., Archibald O. Haller, and Alejandro Portes. 1969. “The Educational and Early Occupational Attainment Process.” *American Sociological Review* 34(1):82–92.

Shen, Ce, and Joseph J. Pedulla. 2000. “The Relationship Between Students’ Achievement and Their Self-Perception of Competence and Rigour of Mathematics and Science: A Cross-National Analysis.” *Assessment in Education* 7(2):237–53.

Steinberg, Laurence. 2008. “A Social Neuroscience Perspective on Adolescent Risk-Taking.” *Developmental Review* 28(1):78–106.

———. 2010. “A Behavioral Scientist Looks at the Science of Adolescent Brain Development.” *Brain and Cognition* 72(1):160–64.

Stinebrickner, Ralph, and Todd R. Stinebrickner. 2014. “A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout.” *Review of Economic Studies* 81(1): 426–72.

Stinebrickner, Todd R., and Ralph Stinebrickner. 2012. “Learning about Academic Ability and the College Drop-Out Decision.” *Journal of Labor Economics* 30(4):707–48.

———. 2013. “Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model.” NBER Working Paper 18945.

Zafar, Basit. 2011. “How Do College Students Form Expectations?” *Journal of Labor Economics* 29(2):301–48.

———. 2013. “College Major Choice and the Gender Gap.” *Journal of Human Resources* 48 (3):545–95.