The High-Stakes Effects of "Low-Stakes" Testing: Performance Labels Matter to Vulnerable Students

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

Harvard Graduate School of Education

December, 2009

Paper prepared for the annual meeting of the American Economic Association, Atlanta, GA, January 3-5, 2010.

*This project was funded by the U.S. Department of Education Institute for Education Sciences (Grant Number R305A080127). 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. Address correspondence to John Papay (john_papay@mail.harvard.edu).

ABSTRACT

Under *No Child Left Behind*, students across the country take annual standardized tests to measure their progress towards state content standards. Although these test results are used to hold schools and districts accountable, the scores on most of the state-mandated tests have no official stakes attached for students. We extend the literature on responses to information by examining whether performance labels students receive on state-mandated tests affect their subsequent educational attainments. A model in which well informed teachers, parents, and students base their actions on the best information available would predict that performance labels would not matter, given that all of these groups have access to the fine-grained test scores on which the performance categories are based. Using a regression-discontinuity design, we test this "no effect of labels" hypothesis by comparing outcomes of students with essentially equal proficiency scoring near the cut score to determine whether being assigned a more positive label improves future outcomes. We find small but substantively important and persistent effects of earning a more positive label on the educational attainments of urban, low-income students. Furthermore, we find that these effects are concentrated among those students who do not plan to attend a four-year college after high school. For these students, simply being classified as "Advanced" rather than "Proficient" on the 10th grade mathematics examination increases by 8 percentage points the probability that they graduate from high school on time and increases by nearly 15 percentage points the probability that they enroll in college.

The High-Stakes Effects of "Low-Stakes" Testing: How Individual Performance Labeling Under *No Child Left Behind* Affects Students

The advent of standards-based reform in American public education has increased dramatically the amount of information available about the mathematics and reading skills of American public school students. Under the *No Child Left Behind* Act of 2001, all states must test students annually in grades 3 through 8 and once in high school. One of the key features of such test-based accountability systems is that every student receives not only a test score but also a performance label (e.g., Failing, Proficient, Advanced) based on their performance. Because the label is assigned entirely based on the test score, in theory it provides no additional information above and beyond the test score itself. Thus, rational economic agents who use all available information should not respond to the label but to the underlying score. However, other models suggest that students, parents, and teachers may not use all available information or that the assignment of a label may affect students' self-concept in ways that affect their subsequent educational success.

In this paper, we use data from Massachusetts to examine the effect of these different performance labels on students' outcomes. The state's practice of assigning these labels to students using well-defined and consistently applied cutoffs presents a natural experiment from which we can draw causal conclusions. Specifically, the state takes the continuous distribution of student performance and divides it into four groups. Students just on either side of any of the three cut scores are assigned to different performance labels despite having essentially equal proficiency. We use a regression discontinuity design to examine whether being assigned a more positive label affects future outcomes for these students on the margin. In all cases, we examine cutoffs where students face no official consequences for their performance – the only benefit of

scoring above the cutoff is earning a more positive performance label. Building on our past work concerning test-based accountability, we focus on urban, low-income students' performances on the mathematics examination across the state of Massachusetts.

The effect we seek to identify is subtle. Students receive both their actual test score and a performance label, so discerning students on the margin should recognize that they were near the cutoff, and falling on one side or the other should not affect their future outcomes. Nonetheless, we find small but substantively important and persistent effects of being classified as "Needs Improvement" instead of "Warning/Failing" or as "Advanced" instead of "Proficient" on important outcomes for urban, low-income students, including their future test scores, their probability of graduating from high school, and their probability of enrolling in college. For example, being labeled "Needs Improvement" rather than "Warning/Failing" on the 8th grade mathematics test increases the probability that students on the margin attend college by two percentage points.

Importantly, we cannot disentangle the mechanisms that underlie these results – students could internalize their performance, affecting them directly, or the mechanism may be more indirect, operating through teachers or parents. We do, however, find that these effects are concentrated among those students who express, on surveys they complete before taking the state tests, that they do not plan to attend a four-year college after high school. For these students, for example, being classified as "Advanced" rather than "Proficient" on the 10th grade mathematics test increases by 8 percentage points the probability that they graduate from high school on time

and by nearly 15 percentage points the probability that they enroll in college. Thus, students with relatively low educational expectations appear to be most affected by the label that they receive.[1]

In the rest of this paper, we describe briefly standards-based reform and examine both the ways in which educational actors use information about student performance in making instructional decisions and the role that performance labels may play in influencing students' subsequent educational outcomes. We then describe our data sources, key measures, and data analytic strategy. 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, arguing that the labeling itself does affect students, either directly by altering their self-concept or indirectly by affecting teachers' or parents' behaviors. These results have important implications for future researchers – particularly those using regression discontinuity designs to estimate causal effects of interventions – and for policymakers and school officials designing and implementing test-based accountability systems.

Background and Context

Standards-based reforms

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 national education policy. One of its major components is the use of standardized testing to monitor student progress toward mastering content standards. Many states, including Massachusetts, implemented test-based accountability programs in the ensuing decades. In 2001, the federal government essentially mandated nationwide testing in its reauthorization of the *Elementary and Secondary Education Act (ESEA)*, called "*No Child Left Behind*" (NCLB). NCLB required all

[1] While the sociological literature distinguishes between "educational aspirations" and "educational expectations," Jacob and Wilder (forthcoming) show that the responses to survey questions aimed at capturing the two concepts are extremely highly correlated. For that reason, we do not distinguish between these concepts.

states taking federal *ESEA* funds to adopt academic standards, to develop an annual testing program to assess student progress toward those standards, and to define what proficient mastery of those standards meant.[2] Currently, states must test all students in grades 3 through 8 and once in high school in mathematics and English language arts, and in several grades in science.

The NCLB legislation set a goal that all American public school students should be proficient by 2014. As a result, all schools must demonstrate annually that they are making "Adequate Yearly Progress" (AYP) toward this goal for students in a variety of specific subgroups, including racial minorities and students with special educational needs or limited English proficiency. This legislation placed consequences on test performance for schools and teachers, with schools that chronically failed to meet AYP subject to increasingly severe sanctions. Although the testing policies under NCLB did not mandate any consequences for students based on their performance, some states have implemented "high stakes" testing for students, particularly in high school. Currently, 26 states, covering nearly three-quarters of the nation's children, 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, 2008).

The use of educational information about student performance

A burgeoning literature suggests that there are significant behavioral responses by educators to scores on examinations to which stakes are attached. For example, because NCLB accountability ratings depend on the number of students who score above a certain proficiency threshold, teachers report pressure to use test score data to identify and focus on "bubble kids," those close to the threshold, rather than students who are likely to score well above or well below

[2] Allowing states to define their own standards, create their own tests, and set their own proficiency levels has produced substantial variation across states in the level of student achievement defined as "proficient."

the cutoff regardless of the amount of attention they receive (Booher-Jennings, 2005). Neal & Schanzenbach (forthcoming) provide quantitative evidence that test-based accountability in Chicago raised test scores for students in the middle of the achievement distribution – near the proficiency cutoff – but not for students at the bottom of the distribution and not consistently for students at the top.

In this paper, we extend the literature on responses to information by examining whether students and teachers respond to labels describing the performances of individual students above and beyond the fine-grained test scores that are available to them. A model in which well informed teachers, parents, and students base their actions on the best information available to them would predict that performance labels would not matter because they provide no information beyond that provided by the students' test score and knowledge of the cut-points used by the state in defining performance categories. However, research in both behavioral economics and cognitive psychology has begun to demonstrate that individuals often act in ways that contradict the predictions of rational actors.

In particular, psychological phenomena can complicate our understanding of human behavior. For example, starting with Brookover, Thomas, and Paterson (1964), sociologists and psychologists have marshaled evidence that students' self-judgments about their potential for academic success can affect educational outcomes (Crocker et al., 2003; Shen & Pedulla, 2000). Claude Steele and Joshua Aronson's work on stereotype threat provides further evidence that individuals' performances on cognitive tasks depend substantially on external factors. Aronson & 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). Specifically, they continue, intellectual competence "is quite literally the product of real or imagined interactions with others.

How a student construes the way he or she is viewed and treated by others matters a lot" (p. 437). Given the importance of quantitative assessments as measures of ability in education today, students' performances on standardized tests are likely important "interactions" that can affect how students view themselves and, thus, can influence their subsequent academic competence.

In other words, we might expect that students who are told that their mathematics performance is "Advanced" might be more likely to work hard in school, more motivated to stay in school, and more likely to think that they are "college material". Even though the student could look at the test score and see that she is right on the border of being simply "Proficient", the specific label she receives might be important. We must note that our analysis cannot differentiate between positive effects of earning a "better" label and negative effects of earning a "worse" label.

Performance labels might also affect students indirectly through the behavioral responses that they produce in teachers and parents. Both parents and teachers may reward or encourage students who score well. For example, a teacher may look through the list of students who pass the test and 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 matter for student outcomes (Jussim & Harber, 2005; Rosenthal & Jacobsen, 1992). In fact, President Bush argued that NCLB was necessary because it challenged the "soft bigotry of low expectations" for disadvantaged and minority students. Importantly, we must note that these indirect effects can take many forms and that they can be either reinforcing or compensatory. In other words, parents may reinforce the positive effect of a student earning the "Advanced" label by rewarding them or beginning to talk to them about college, or they may compensate for students who do not earn

this label by providing additional tutoring or supports. Obviously, these responses operate in different directions, and we cannot disentangle them here.

Research Questions

In this paper, we examine whether the performance labels that students receive under the Massachusetts test-based accountability system affect their subsequent educational outcomes. We focus on test classifications that have no official consequences for students. In particular, we examine the effects of being classified as "Needs Improvement" instead of "Warning/Failing" on the 8th grade test and as "Advanced" instead of "Proficient" on the 8th and 10th grade mathematics tests. Given that our previous work in Massachusetts has suggested that urban, low-income students are much more sensitive to the effects of test performance labeling than their suburban or wealthier peers on high-stakes examinations, we retain our focus on this traditionally disadvantaged group that has fared poorly in American public education.

We might expect that students' perceptions of their educational futures might moderate the effects of performance labeling. For example, students who feel confident about their academic abilities might not be greatly affected by the label they receive, but students who are more vulnerable might experience a greater benefit from earning a positive classification. To examine whether performance labels affect educational outcomes more for some students than for others, we make use of students' responses to survey questions about their post-secondary educational plans.

The research community has long known that student aspirations predict educational attainment (Duncan, Featherman, & Duncan, 1972; Sewell, Haller, & Ohlendorf, 1970; Sewell, Haller, & Portes, 1969). However, the development of educational aspirations is a process that researchers do not fully understand. A recent review of the research evidence suggests that

students' aspirations depend on their academic abilities and motivation, their parents' educational attainments and family income, and the attitudes and aspirations of their peer group, among other things (Jacob & Wilder, forthcoming). We use these data about students' aspirations in two ways. First, we examine the extent to which performance labels affect students' plans in the future. Second, we examine whether the effects depend on the students' initial post-secondary educational plans. In other words, we examine whether students' aspirations moderate the impact of MCAS failure on subsequent educational attainments.

To summarize, our specific research questions are:

RQ1. Does earning a more positive performance label on the Massachusetts state mathematics test affect future educational aspirations, test scores, and attainments for urban, low-income students on the margin of passing?

RQ2: Do students' educational aspirations moderate these effects?

Research Design

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) tests in 1998. For most students, performance on these tests carries no consequences; however, starting with the class of 2003, the 10th grade tests became high-stakes exit examinations that students must pass to graduate.

This focus on standards-based reform and these investments in public education appear to have paid off. The state has been praised for having the most rigorous academic standards in the country and high-stakes examinations that align closely with these standards (Finn, Julian, & Petrilli, 2006; *Quality Counts*, 2006). Furthermore, Massachusetts students are consistently among the nation's top performers on the National Assessment of Educational Progress (NAEP) examinations, and the state's NAEP performance has improved rapidly since the introduction of state testing (NCES, 2008). However, the state still faces significant educational challenges, including large gaps between the average achievement of students of color and that of non-Hispanic white students (MA DOE, 2009).

To address our research questions, we have synthesized several datasets 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, high school graduation status, and responses to surveys that students complete just before taking the MCAS examinations. We have supplemented this dataset with several sources of information, including records from the National Student Clearinghouse. Thus, our dataset includes information about college attendance, including the date of enrollment, institution attended, and institutional characteristics. As a result, we can measure students' initial post-secondary educational attainments.

We focus on testing in 8th and 10th grade mathematics. On the 10th grade examination, which is a high-stakes test that students must pass to graduate from high school, we focus on students who fall well above the passing cutoff. We choose the mathematics examination because our past work in this area found that, for urban, low-income students on the margin of

passing, barely passing the 10th grade mathematics examination increased the probability of graduation substantially, but there were no effects of performance on the ELA examination (Papay, Murnane, & Willett, forthcoming).

Students receive information about their performance in detailed reports. Appendix A includes an example of one such report; it provides students with information about their test score, a range of "likely values" the student would earn if she took the test several times, and a performance label. It also contains a range of interpretive information (not shown) to help students and parents make sense of their test scores. Thus, students and parents receive a wide range of information about their performance.

We pool data across several years, examining the group of students who took the 8th and 10th grade mathematics examinations in the spring of 2003 through 2006. These students are members of the graduating cohorts of 2005 through 2009. For each year, we restrict our sample to students who took the MCAS examination for the first time in that grade.

Measures

To address our research questions, we created four outcome variables. Two focus on students' educational attainments: a dichotomous outcome variable that indicates whether the student graduated from high school on-time with their cohort (*GRAD*) and a dichotomous outcome variable that indicates whether the student attended college within one year of cohort graduation (*COLL*) using data from the *National Student Clearinghouse*. We defined on-time cohort graduation as occurring four years after the student took the 8th grade examination and two years after the students took the 10th grade examination. We counted a student as attending college if they were recorded as having been enrolled in college by June 1 one year after their cohort graduation.

For our analysis of the 8th grade test, we use two additional outcomes. The first is the student's 10th grade MCAS mathematics test score (*MATH_GR10*). The second concerns students' educational aspirations and comes from a survey that students complete immediately before they take the 10th grade MCAS examination. Of particular interest, this survey asks students about their post-secondary educational plans. Although the state has made minor changes to the question over the years, all versions are quite similar to the one from the 2005 administration:

> 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 10th graders since the 2002-03 school year and of all 8th graders starting in 2005-06. Eighty-three percent of Massachusetts 10th grade students completed this survey. This sample is not fully representative of all Massachusetts students; for example, low-income and urban students are somewhat less likely to have completed it. Thus, our results for the subsample of students who took the survey cannot generalize to the full population of Massachusetts test-takers; given our identification strategy described below, however, this limitation does not affect the internal validity of our study. We focus on whether students plan to attend a four-year college or not. We code a dichotomous predictor (COLL_ASP) to indicate whether the student reported that they planned to attend a four-year college. In Massachusetts, 69% of all students who completed the 10th grade survey – and 62% of low-income urban students – reported planning to attend a four-year college.

The extent to which we can examine each of these four primary outcomes depends on the timing of the initial test and outcome data collection. In other words, we must have five years of data after the 8th grade test to examine the effect of classification on college outcomes, but only two years to examine the effects on 10th grade mathematics scores. As seen in Table 1, each of our analyses uses a different number of years of data. Importantly, since survey responses of 8th graders to the question about post-secondary educational plans are only available for the 2005-06 cohort, the analyses that examine our second research question for 8th graders use only one year of available data and cannot examine high school graduation or college attendance as outcomes.

INSERT TABLE 1 ABOUT HERE

Our key predictors come from the state testing dataset, which contains a record of scores from every MCAS examination that each student took from 3rd grade through high school graduation. The state reports test information at four levels: test item information, raw scores, scaled scores, and performance level. The state uses a 3-parameter item-response theory (IRT) scaling model to generate the scaled scores, which range from 200 to 280 in increments of two points. A score of 220 qualifies as passing, with a different performance rating each 20 points, as follows:

(a) 200 to 218: Failing/Warning
(b) 220 to 238: Needs Improvement
(c) 240 to 258: Proficient
(d) 260 to 280: Advanced

Because the scaled scores have such a coarse scale, with multiple raw scores translating to a single scaled score, we use raw scores in our analyses.[3]

To implement our regression-discontinuity approach, we center students' raw scores by subtracting out the value of the corresponding minimum passing score. Because the state uses a

[3] For more information on MCAS scoring and scaling, see the MCAS Technical Reports (MA DOE, 2002, 2005).

complicated scaling procedure that depends on the sample of students who take the test each year, this minimum passing score changes from year to year.[4] On this re-centered continuous predictor (*MATH*), a student with a score of zero had achieved the minimum passing score. We also created a dichotomous version of this same predictor (*ABOVE*) to indicate whether the student fell above the cut score or not (coded 1 if the student scored at or above the relevant cut score; 0 otherwise).[5]

We also include selected control predictors in our analyses, including: (a) dichotomous predictors that describe student race, gender, whether the student was limited English proficient, a migrant, an immigrant, required special education, and/or enrolled in Career and Technical education, (b) student's ELA test performance, and (c) the fixed effect of cohort. We focus our analyses on students who are eligible for federal free or reduced price lunch programs who are enrolled in one of Massachusetts's 22 urban school districts.[6] Overall, 26% of Massachusetts 10th grade students (and 28% of 8th graders) attended urban schools and 31% of 10th graders (29% of 8th graders) were identified as low income. Low-income students tended to cluster in the urban schools: in 10th grade, 68% of urban students lived in poverty, compared to just 18% of suburban students.

Data Analyses

Our analyses use the regression-discontinuity design.[7] Introduced in the late 1960s, this research strategy has grown in popularity because it is one of the most rigorous methods for drawing causal inferences outside of a true, randomized experiment (Cook, 2008; Shadish, Cook, & Campbell, 2002). Conceptually, we would like to take students who scored identically, right at

[4] For example, in 2004, students need to earn 21 of a possible 60 points to pass the mathematics examination.
[5] 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.
[6] The state defines urban districts as those that participate in the state's Urban Superintendents Network.
[7] This paragraph draws heavily from Papay, Murnane, & Willett, forthcoming.

each cut score, and randomly assign them to a performance label. This assignment process would render them equal in expectation on all observable and unobservable characteristics prior to treatment, allowing us to identify any differences in the ultimate outcome as a causal effect of being assigned the performance label rather than of earning higher scores.

Of course, we cannot conduct this experiment on the state test; however, the state's exogenous imposition of cutoffs provides a natural experiment from which we can draw equivalent causal conclusions. By examining students near each cut score, we can extrapolate outcomes for two groups – those who scored at the exogenously-assigned cut score and were assigned the more positive label (represented by parameter γ_{above}) and those students who would have scored at the cut score and received the less positive label (represented by parameter γ_{below})[8]. The difference between these two parameters provides an unbiased estimate of the causal impact of the classification for students at the cut score. Thus, we obtain an estimate of the average treatment effect for students "on the margin" of the cutoff.

To address our first research question, we use a regression-discontinuity design to examine whether barely earning a more positive label on the examination affects outcomes for students on the margin. We focus here on our analysis of our high school graduation outcome, but we use the same analytic strategy for all of our outcomes. In its basic formulation, this approach involves fitting a linear probability model of the following form:

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

for the ith student. In this model, β_2 represents the causal effect of interest. If its estimated value is statistically significant and positive, then we can conclude that classifying a student at the cut score as earning the more positive label, as opposed to earning the less positive label, causes the

[8] Technically, $\gamma_{above} = \lim_{MATH_i \to 0^+} [P(GRAD_i = 1) \mid MATH_i]$ and $\gamma_{below} = \lim_{MATH_i \to 0^-} [P(GRAD_i = 1) \mid MATH_i]$

student's probability of attending college to *increase* discontinuously.

The internal validity of our regression-discontinuity analyses – and consequently our ability to make unbiased causal inferences – rests on two key assumptions. First, the state must apply the cut score consistently such that every student who falls below it earns the less positive performance label and students cannot manipulate their position relative to the cut score. Second, we assume that we can model credibly the underlying relationship between student MCAS score and the probability of graduation. In equation (1), we present this relationship as linear. However, because we do not know the exact functional form of this relationship, we relax this assumption. Because our parameters of interest – γ_{above} and γ_{below} – represent limits estimated at a boundary point, we use the nonparametric smoothing method of local linear regression recommended by Hahn, Todd, & Van der Klaauw (2001).[9]

Our implementation of this strategy follows the approach laid out by Imbens and Lemieux (2008). We select a bandwidth (h) to govern the amount of smoothing in the local linear regression analysis. We choose an optimal bandwidth (h*) for each analysis using a well-defined statistical fit criterion and a cross-validation procedure described by Imbens & Lemieux (2008).[10] We slide this bandwidth smoothly through the data, fitting locally linear regression models at each score point. Connecting these regressions creates the nonparametrically smoothed fit we present in our figures.

We then estimate our parameter of interest, the difference between γ_{above} and γ_{below}, in one step, by fitting the model presented in equation (1) using observations that fall only within

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

[10] $h^* = \arg\min_{h} \frac{1}{N}\sum_{i=1}^{N}(G\hat{R}AD_i(h) - GRAD_i)^2$, where $G\hat{R}AD_i(h)$ 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.

one bandwidth on either side of the relevant cut score.[11] We extend this approach in several ways. First, we add a vector of selected student background covariates to the model in (1) to improve the precision of our estimation. We also include the fixed effect of cohort to account for average differences in our outcome across years.[12] Since we have a strong prior belief that obtaining a more positive performance label would not reduce educational outcomes, we use one-tailed tests.

To address our second research question, we fit a statistical model similar to that specified in (1). In this case, we include the student's pre-treatment self-reported college aspirations ($COLL_ASP_i$) as a covariate in our model and interact it with the main predictors, $ABOVE_i$ and $MATH_i$, as follows:

$$p(GRAD_i = 1) = \beta_0 + \beta_1 MATH_i + \beta_2 ABOVE_i + \beta_3 (ABOVE_i \times MATH_i) + \beta_4 (MATH_i \times COLL_ASP_i) \qquad (2)$$
$$+ \beta_5 (ABOVE_i \times COLL_ASP_i) + \beta_6 (ABOVE_i \times MATH_i \times COLL_ASP_i) + \beta_7 COLL_ASP_i + \varepsilon_i$$

for the ith individual within one bandwidth on either side of the relevant cut score. In this model, β_2 represents the causal effect of receiving the more positive performance label on the population probability of on-time high school graduation for students at the margin of passing who *did not* plan to attend a four-year college. The parameter sum $\beta_2 + \beta_5$ represents the causal effect of earning a positive label for students at the margin of passing who *did* plan to attend a four-year college. For these analyses, we necessarily restrict our sample to low-income urban students who completed the survey.

Findings

[11] 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.
[12] We tested whether adding school fixed effects would significantly increase the explanatory power of our models, and found that they did not. We also found that the critical coefficients were not sensitive to the decision of whether to include school fixed effects in the model.

Research Question 1: Does earning a more positive performance label on the Massachusetts state mathematics test affect future educational aspirations, test scores, and attainments for urban, low-income students on the margin of passing?

We find strong evidence that earning a more positive performance label affects future outcomes for urban, low-income students. The effects are small, but substantively important. Again, we examine the effects of earning performance labels at three different cutoffs: the 8th grade Needs Improvement/Warning cutoff, the 8th grade Advanced/Proficient cutoff, and the 10th grade Advanced/Proficient cutoff. In each case, students do not face any consequences for their performance and get no official benefit from earning the more positive label. As a result, any differences we find among students with essentially equal proficiency scoring near these cutoffs represent the effects of the label itself.

As seen in Table 2, we find that being classified as "Needs Improvement" as opposed to "Warning" in 8th grade increases students' 10th grade MCAS mathematics scores by one-third of a point (p=.03), or approximately 0.04 of a standard deviation.[13] Furthermore, earning the more positive label increases the probability of graduating from high school on-time by 3.2 percentage points (p=.003) and of enrolling in college by 2.3 percentage points (p=.007).[14] These effects are persistent and substantively large. For example, given that only 64% percent of urban, low income students scoring near the margin graduate on-time, this 3.2 percentage point difference represents a substantial effect.

INSERT TABLE 2 ABOUT HERE

The evidence is not as consistent that being classified as "Advanced" instead of "Proficient" affects student outcomes. Although scoring "Advanced" increases students' 10th

[13] Calculated among students with similar proficiency on the 8th grade examination.
[14] Again, all p-values are derived from one-tailed tests.

grade MCAS scores by 0.39 points (p=.036), these effects do not appear to persist. As seen in Table 2, there is no effect of the classification on graduation and the estimate of a 1.9 percentage point effect on college attendance does not reach levels of traditional statistical significance (p=0.112).

However, we do find that classification as "Advanced" on the 10th grade examination improves student outcomes. Earning the "Advanced" rating increases by 2.4 percentage points (p=0.011) student's probability of graduating from high school and by 2.3 percentage points (p=0.002) the probability that they enroll in college. Importantly, the 10th grade "Advanced" cutoff does have consequences for some students – those who are eligible for the state-sponsored Adams Scholarship for college. Students are eligible for this scholarship if they score Advanced in either mathematics or ELA and Proficient in the other. As a result, for some students, earning the Advanced designation could make them eligible for the scholarship. Thus, we conduct a secondary analysis in which we exclude students who scored Proficient in ELA – the only students for whom the Advanced designation would matter for the scholarship. Our results are quite similar to those presented above.

Thus, we conclude that urban low-income students – or their parents or teachers – do indeed respond to the performance label they receive, not simply the information contained in their test score. Among students with equal proficiency near a cut score, earning the more positive label increases later educational outcomes; these improvements are substantively important.

RQ2: Do students' educational aspirations moderate these effects?

Our hypothesis that at least some of the effects of performance labels operates through students' conceptions of their own ability led us to consider heterogeneity within this urban, low-

income group. We might expect that for some students – or their parents – test performance does not affect their ideas about their abilities, while other students are more susceptible to the effects of performance labeling. As a result, we examined two groups of students – those who reported that they planned to attend a four-year college after high school and those who reported having other plans. We find quite strong evidence that educational aspirations moderate the effects of classification on these tests. For students who do not plan to attend a four-year college, earning a more positive label has a substantial, positive effect across all cutoffs and outcomes,

Importantly, our analysis of the 8th grade results is limited by data availability. The state first gave the 8th grade survey to students in 2006, so we cannot examine important educational attainment outcomes like high school graduation or college attendance. Instead, we must rely on proxies, such as students' 10th grade test scores and their expressed educational aspirations in 10th grade. Nonetheless, the results are striking. As seen in Table 3, being classified as "Needs Improvement" instead of "Warning/Failing" on the 8th grade mathematics examination increases 10th grade MCAS scores by more than 1 point (p=0.008) – or 0.14 of a standard deviation – for urban low-income students who do not plan to attend college. The more positive label also raises students' probability of expressing four-year college as their intended post-secondary goal on the 10th grade survey by 12 percentage points (p=.011). These effects are strikingly large. There are no effects for students who plan to attend college.

INSERT TABLE 3 ABOUT HERE

Very similar patterns appear at the Advanced/Proficient cutoff in 8th grade, although again the point estimates are at the margin of statistical significance. Earning the "Advanced" label increases students' 10th grade mathematics test scores by nearly three points (p=0.009) and

raises the probability that students will express four-year college aspirations in 10th grade by 15 percentage points (p=0.053).

Again, we find similar patterns on the 10th grade examination, a test on which we can also measure educational attainment outcomes of interest. Performance labeling appears to be particularly important for students who do not plan to attend a four-year college after graduation. For these students, being classified as "Advanced", rather than "Proficient", increases the probability that they will graduate from high school on-time by 8.4 percentage points (p=0.043) and that they will attend college by 14.8 percentage points (p=0.002).

This second result is particularly striking – the fitted probability of attending college increases from 34% to 48% for these students simply by being labeled "Advanced" instead of "Proficient". In Figure 1, we present the fitted nonparametrically smoothed relationship between college attendance and raw MCAS mathematics scores for urban, low-income students with and without four-year college aspirations. This figure reveals several important lessons. First, students who report as 10th graders that they plan to attend a four-year college do indeed attend college at a substantially greater rate than students with other plans; the vertical distance between the two lines is substantial. Second, there is no disruption in the smoothed relationship for students with college aspirations. In other words, being classified as "Advanced" does not affect their probability of attending college. However, for students without four-year college aspirations, earning the more positive performance label substantially increases the probability of attending college. This effect is seen in the sharp and substantial disruption in the probability of college attendance at the cut score.

INSERT FIGURE 1 ABOUT HERE

Threats to Validity

The internal validity of a regression discontinuity design depends on two important assumptions. First, treatment assignment – here the performance label students scoring near the cutoff receive – must be exogenous and applied rigidly to all students. All student characteristics, both observed and unobserved, must differ smoothly around the cut score. In other words, students must not be able to manipulate their position relative to the cut score. This assumption appears to hold as all students who score below the cut-off are assigned one label and all students who score above the cut-off are assigned a more positive label. Furthermore, the cut scores vary from year-to-year based on a complicated scaling formula and are determined after students take the test; thus, the probability that a student could intentionally score just above a cut-off is quite small.

We assess whether the cut scores were imposed exogenously in several ways. First, we examine the density of students falling on either side of the cut score; in Figure 2, we present the distribution of MCAS mathematics scores for students on the grade 10 examination near the "Advanced"/"Proficient" cutoff. There is no discontinuity in this density at the cut score. Second, we examine whether there are apparent discontinuities at the cut score in student-level covariates. We find no reason to doubt that the state has imposed the cut score exogenously and consistently.

INSERT FIGURE 2 ABOUT HERE

The second key assumption underpinning our regression-discontinuity analyses is that we can estimate the relationship between the outcome and the test score accurately, at least in the immediate vicinity of the cut score. In part, we address this issue by using a flexible, local linear regression approach to model this relationship. The key decision in this analysis then becomes the choice of bandwidth, h, that governs the amount of smoothing. To assess the sensitivity of

our results to this decision, we refit our principal models restricting the sample to students whose test scores fall within different bandwidths around the cut-off.

In Table 4, we present the results from this analysis. We find that our main findings are quite robust to the choice of bandwidth. For example, being classified as "Needs Improvement" rather than "Warning/Failing" on the 8th grade examination increases the probability that students attend college by between 2.3 and 4.3 percentage points, depending on bandwidth. Some individual estimates are less robust; for example, the effect of being classified as "Advanced" rather than "Proficient" in 8th grade on 10th grade MCAS scores is sensitive to bandwidth. However, the general pattern that a more positive performance label produces better student outcomes persists across most analyses. In particular, the effects of earning a more positive classification for students who do not plan to attend a four-year college are consistently large, positive, and statistically significant.

INSERT TABLE 4 ABOUT HERE

One threat to all empirical studies that present results for several outcomes and for sub-groups is that they are subject to Type I error. Of course, our research is not immune to this concern. However, two patterns in our results buttress the claim that our findings reflect causal impacts rather than Type I errors. First, the findings are consistently stronger across all outcomes for the group of low-income students who reported low educational aspirations than for those who reported higher aspirations. Second, we examined whether the patterns we find are consistent across years. We found that they were, although the power of our tests is substantially reduced by the limited sample size. For example, our estimates of the effect of earning an "Advanced" label in 10th grade on college attendance for students without college aspirations are 13.4 percentage points in 2004, 10.8 percentage points in 2005, and 15.6 percentage points in

2006. In all cases, earning the more positive label has a substantial effect on student outcomes.

Finally, one key limitation of the regression-discontinuity approach is that we must focus only on students near the margin of passing, thereby restricting the external validity of our study. Our analysis is an example of using the regression-discontinuity design to examine different points in the distribution. As the state sets multiple cut scores, we can examine these effects at different cut scores. Interestingly, we find relatively similar patterns at each level, although we do see some evidence that different margins may be at play at different points in the distribution.

Discussion

In this paper, we examine the extent to which individuals respond to the performance labels students receive on state tests. Given that students, parents, and teachers receive the test score and the performance label, the label itself in theory adds no additional information. However, we find strong evidence that performance labeling affects future student educational outcomes and attainments. In particular, earning a more positive label – even on low-stakes tests that carry no official consequences for students – produces a substantial and persistent benefit for urban, low-income students on the margin of passing. These effects are more pronounced for students who do not plan to attend a four-year college after graduation. For example, on the 10th grade test, being classified as "Advanced", rather than "Proficient", increases the probability that these students will graduate on-time from high school by 8 percentage points and that they will attend college by nearly 15 percentage points.

The interpretation of these findings proves challenging. First, we do not know whether they reflect the positive effects of earning a "better" label or the negative effects of earning a "worse" label. In other words, students who are labeled as "Advanced" could be encouraged by their performance, which could lead them to continue in school or to be more motivated in their

courses. However, relatively high performing students who are labeled as simply "Proficient" may be discouraged by their failure to achieve the more prestigious label and may be more likely to consider themselves not "college material". Unfortunately, since each group represents our estimate of the counterfactual for the other, we can only examine the difference between them. Regardless, though, students with equal proficiency have different outcomes based simply on the label that they receive.

Furthermore, although we find that the label itself matters, we cannot precisely determine the mechanisms through which these effects operate. For example, students could respond directly, feeling encouraged or discouraged as a result of their performance. However, parents or teachers may also respond, producing indirect effects on student outcomes. For example, teachers may simply examine a list of students scoring at each performance level, explicitly using only the information contained in the labels. Thus, their attitudes and expectations about students may be formed by these test labels and may in turn affect their interactions with students. Our results do suggest that at least part of the effect operates through student – rather than teacher – responses. If teacher responses were driving these patterns, then we should not expect to see differences among students in the same schools based on their college aspirations. However, we find that effects are concentrated much more heavily in students who do not express four-year college plans, which suggests that the students' attitudes are important in governing these behavioral responses.

This paper raises several implications for applied econometric research, particularly using the regression-discontinuity design. First, this paper illustrates the importance of thinking carefully about heterogeneity of causal effects. We find clear evidence that the effect of performance labeling is greatest among students who do not plan to attend a four-year college.

Presumably this measure of educational aspirations reflects a constellation of individual (and family) characteristics, part of which involves the students' own attitudes about their abilities. Educational research has begun to examine effect heterogeneity along traditional demographic characteristics, but our results suggest another important dimension of student profiles, one that is not well predicted by demographic characteristics in our sample of low-income urban students.

Finally, and most importantly, the fact that we find behavioral responses associated with particular performance labels calls into question other identification strategies that seek to use similar discontinuities as exogenous variation to identify the causal effects of other interventions. In such an analysis, individuals are typically assigned to a particular intervention based on their position relative to the cut score and researchers can exploit this cutoff-based assignment using a regression-discontinuity design. However, if individuals respond to the label itself, and not just to the intervention, estimates about the intervention's effects may be confounded with the effect of labeling itself.

For example, in earlier work we examined the effects of barely failing the Massachusetts high school exit examination on student educational attainments (Papay, Murnane, & Willett, forthcoming). Other researchers have examined similar questions in Texas, California, and New Jersey (Martorell, 2004; Reardon et al., 2009; Ou, 2009). In all cases, students who fail the examination must retake and pass it before they can graduate from high school. Thus, failing raises a structural barrier to graduation – the need to retake the test – that may cause students to drop out of school. However, our work here suggests that the effect may also operate through more complicated mechanisms and that simply interpreting the effect of "failing" as a consequence of the exit examination requirement may be an overstatement. The results in this paper complicate our understanding of this research concerning exit examinations.

As another example, Jacob and Lefgren (2004) examine the effect of summer school on student outcomes. Students in Chicago were assigned to summer school largely on the basis of their performance on the district's accountability test; students who failed the test were assigned to summer school, while students who passed did not need to attend. Because the summer school attendance rule was not adhered to precisely, the authors used a fuzzy regression-discontinuity design. They concluded that summer schooling had positive effects on students. However, to the extent that students' placement relative to the summer schooling cutoff constituted a performance label that mattered to students, we may see some benefit of passing the test even if summer school had not been a consequence of failing. As a result, the effects of summer schooling that Jacob and Lefgren identify may be understated if there were also positive effects of simply being labeled as "passing" the test.

In short, this work suggests that psychological or other mechanisms 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.

This paper also raises important substantive implications. We find clear evidence that individual actors are not using the full range of information available to them. Given that the state has invested in providing parents and students with detailed and clear reports concerning student performance, this result is particularly interesting. It appears that, on average, urban low-income students (or their parents or teachers) use the information contained in the performance label itself, even though finer-grained information about test performance is available. This is

particularly true for students who do not plan to attend a four-year college after high school. This group appears to be particularly vulnerable to the effects of labeling, suggesting that these students are especially sensitive either to positive encouragement or negative reinforcement of their attitudes.

In order to formulate policy responses to the evidence that performance labels matter, it is important to learn whose behaviors are responding to the labels and the mechanisms through which the labels affect subsequent educational outcomes. We plan to explore these questions in subsequent papers.

References

Aaronson, J. and Steele, C. (2005). "Stereotypes and the fragility of academic competence, motivation, and self-concept." In A.J. Elliot and C.S. Dweck, eds., *Handbook of competence and motivation.* New York: Guilford Press.

Booher-Jennings, J. 2005. "Below the Bubble: "Educational Triage" and the Texas Accountability System" *American Educational Research Journal*, *42*(2): 231-268.

Brookover, W.B., Thomas, S., & Paterson, A. (1964). Self-concept of ability and school achievement. *Sociology of Education, 37*:271-278.

Center on Education Policy. (2008). *State high school exit exams: Moving toward end-of-course exams*. Retrieved November 15, 2008, from http://www.cep-dc.org/document/docWindow.cfm?fuseaction=document.viewDocument&documentid=244&documentFormatId=3803.

Cook, T.D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics, and economics. *Journal of Econometrics, 142*(2): 636-654.

Crocker, J., Karpinski, A., Quinn, D.M., & Chase, S.K. (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), 507–516

Duncan, O., Featherman, D., and Duncan, B. (1972). *Socioeconomic background and achievement*. New York: Seminar Press.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association,* 87(420), 998-1004.

Finn, C.E., L. Julian, & 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.

Hahn, J., P. Todd, & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica,* 69(1), 201-209.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-35.

Jacob, B.A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, 86*(1): 226-244.

Jacob, B.A., & Wilder, T. (forthcoming). Educational expectations and attainment. Paper prepared for the Social Inequality and Educational Disadvantage conference, Washington, D.C.

Jussim, L. and Harber, KD (2005) "Teacher Expectations and Self-Fulfilling Prophecies: Knows and Unknowns, Resolved and Unresolved Controversies." Personality and Social Psychology Review, 9(2), 131-155.

Lee, D.S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics, 142*(2), 655-74.

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

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 Elementary and Secondary Education. (2009). *Spring 2009 MCAS tests: Summary of state results.* Retrieved December 26, 2009, from http://www.doe.mass.edu/mcas/2009/results/summary.pdf.

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. (forthcoming). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics.*

Ou, D. (2009). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *CEP Discussion Paper No 907*. Retrieved February 24, 2009 from http://cep.lse.ac.uk/pubs/download/dp0907.pdf.

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

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

Reardon, S.F., Arshan, N., Atteberry, A., & Kurlaender, M. (2008). High stakes, no effects: Effects of failing the California High School Exit Exam. Paper prepared for the Annual Meeting of the Association for Public Policy Analysis and Management, Los Angeles.

Rosenthal R., & Jacobson, L. (1992). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development.* New York: Irvington.

Sewell, W. H., Haller, A. O., and Ohlendorf, G. W. (1970). The educational and early occupational status attainment process: Replication and revision. *American Sociological Review* 35, 1014-27.

Sewell, W. H., Haller, A. O., and Portes, A. (1969). The educational and early occupational attainment process. *American Sociological Review*, 34, 82-92.

Shadish, W.R., T.D. Cook, & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Company.

Shen, C. & Pedulla, J.J. (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), 237-253.

Table 1. Description of data structure and years available for analysis of specific predictors and outcomes.

8th Grade Analysis			
8th Grade Test Cohort	10th Grade MCAS	On-time high school graduation	College attendance
2002-03	2004-05	2006-07	2007-08
2003-04	2005-06	2007-08	2008-09
2004-05	2006-07	2008-09	--
2005-06	2007-08	--	--
10th Grade Analysis			
10th Grade Test Cohort	10th Grade MCAS	On-time high school graduation	College attendance
2002-03	--	2004-05	2005-06
2003-04	--	2005-06	2006-07
2004-05	--	2006-07	2007-08
2005-06	--	2007-08	2008-09
2006-07	--	2008-09	--

Table 2. 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. Cell entries include the parameter estimate, standard error (in parentheses), optimal bandwidth used, sample size, and asterisks to denote inference.

	8th Grade		10th Grade
	Needs Improvement/ Warning	Advanced/Proficient	Advanced/Proficient
Grade 10 MCAS Score	0.333 *	0.389 *	
	(0.144)	(0.194)	N/A
	h=3	h=5	
	9,082	3,282	
On-time Graduation	0.032 **	-0.011	0.024 *
	(0.009)	(0.023)	(0.009)
	h=5	h=8	h=7
	14,307	4,086	9,050
College Attendance	0.023 **	0.019	0.023 **
	(0.007)	(0.015)	(0.007)
	h=3	h=8	h=6
	4,631	2,259	7,816

NOTE: *, p<0.05; **, p<0.01; ***, p<0.001. All p-values are derived from one-tailed tests.

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, by whether they express aspirations to attend a four-year college after high school. Cell entries include parameter estimates, standard errors (in parentheses), and asterisks to denote inference.

	Students with College Aspirations	Students Without College Aspirations	Sample Size
8th Grade Needs Improvement/Warning Cutoff			
10th Grade MCAS Score	0.572	1.315 **	1,826
	(0.537)	(0.398)	h=3
Express College Aspirations (Grade 10)	0.005	0.12 *	2,143
	(0.022)	(0.044)	h=5
8th Grade Advanced/Proficient Cutoff			
10th Grade MCAS Score	0.237	2.903 **	938
	(0.332)	(1.020)	h=5
Express College Aspirations (Grade 10)	0.05 *	0.153	901
	(0.025)	(0.087)	h=6
10th Grade Advanced/Proficient Cutoff			
On-time Graduation	-0.015	0.084 *	6,048
	(0.017)	(0.046)	h=7
College Attendance	0.004	0.148 **	5,224
	(0.015)	(0.041)	h=6

NOTE: *, p<0.05; **, p<0.01; ***, p<0.001. All p-values are derived from one-tailed tests.

Table 4. Estimated effect of earning the more positive performance label at different cutoffs and on different outcomes, by bandwidth.

	Bandwidth				
	h*-2	h*-1	h*	h*+1	h*+2
8th Grade Needs Improvement/Warning cutoff - All Urban, Low-Income Students					
10th Grade MCAS Score	--	0.728 **	0.333 *	0.32 *	0.055
		(0.108)	(0.144)	(0.139)	(0.235)
On-time Graduation	0.024 *	0.031 **	0.032 **	0.016	0.007
	(0.008)	(0.010)	(0.009)	(0.012)	(0.011)
College Attendance	--	0.028 ***	0.023 **	0.036 **	0.043 **
		(0.002)	(0.007)	(0.012)	(0.012)
8th Grade Needs Improvement/Warning cutoff - Urban, Low-Income Students without Four-Year College Aspirations					
10th Grade MCAS Score	--	0.994 *	1.315 **	0.944 *	0.413
		(0.361)	(0.398)	(0.352)	(0.383)
Express College Aspirations (Grade 10)	0.144 *	0.105	0.12 *	0.097 *	0.049
	(0.052)	(0.057)	(0.044)	(0.046)	(0.053)
8th Grade Advanced/Proficient cutoff - All Urban, Low-Income Students					
10th Grade MCAS Score	0.746 ***	0.65 ***	0.389 *	0.007	-0.155
	(0.095)	(0.124)	(0.194)	(0.341)	(0.328)
On-time Graduation	-0.014	-0.026	-0.011	-0.012	-0.003
	(0.023)	(0.023)	(0.023)	(0.021)	(0.021)
College Attendance	0.005	0.002	0.019	0.021	0.026 *
	(0.013)	(0.013)	(0.015)	(0.013)	(0.013)
8th Grade Advanced/Proficient cutoff - Urban, Low-Income Students without Four-Year College Aspirations					
10th Grade MCAS Score	4.274 **	4.044 ***	2.903 **	1.599	1.632
	(0.964)	(0.602)	(1.02)	(1.221)	(1.000)
Express College Aspirations (Grade 10)	0.158	0.153	0.153	0.164 *	0.178 *
	(0.116)	(0.093)	(0.087)	(0.085)	(0.084)
10th Grade Advanced/Proficient cutoff – All Urban, Low-Income Students					
On-time Graduation	0.029 **	0.022 *	0.024 *	0.016	0.017 *
	(0.011)	(0.010)	(0.009)	(0.010)	(0.009)
College Attendance	0.035 ***	0.032 ***	0.023 **	0.036 ***	0.03 ***
	(0.007)	(0.007)	(0.007)	(0.009)	(0.007)
10th Grade Advanced/Proficient cutoff - Urban, Low-Income Students without Four-Year College Aspirations					
On-time Graduation	0.084	0.102 *	0.084 *	0.055	0.063
	(0.047)	(0.048)	(0.046)	(0.045)	(0.044)
College Attendance	0.19 **	0.169 **	0.148 **	0.146 **	0.119 **
	(0.056)	(0.046)	(0.041)	(0.039)	(0.040)

Figure 1.

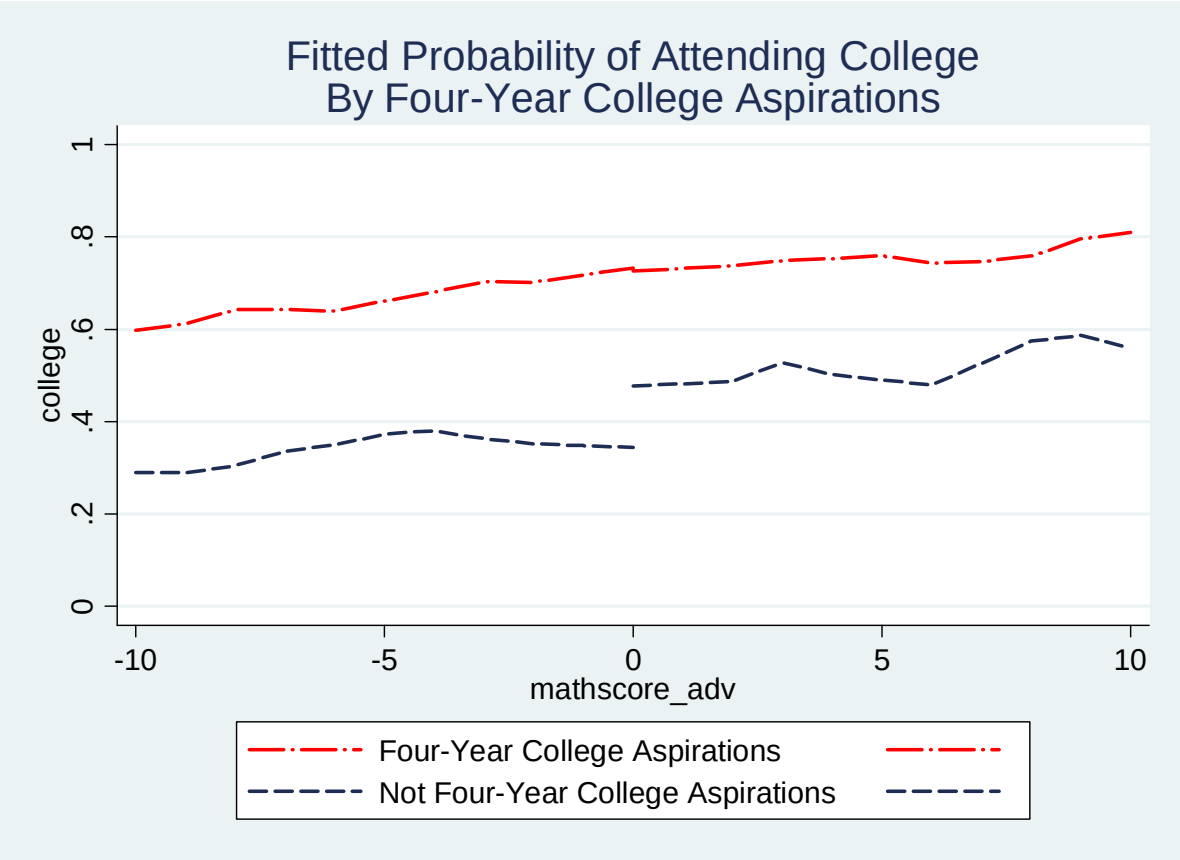
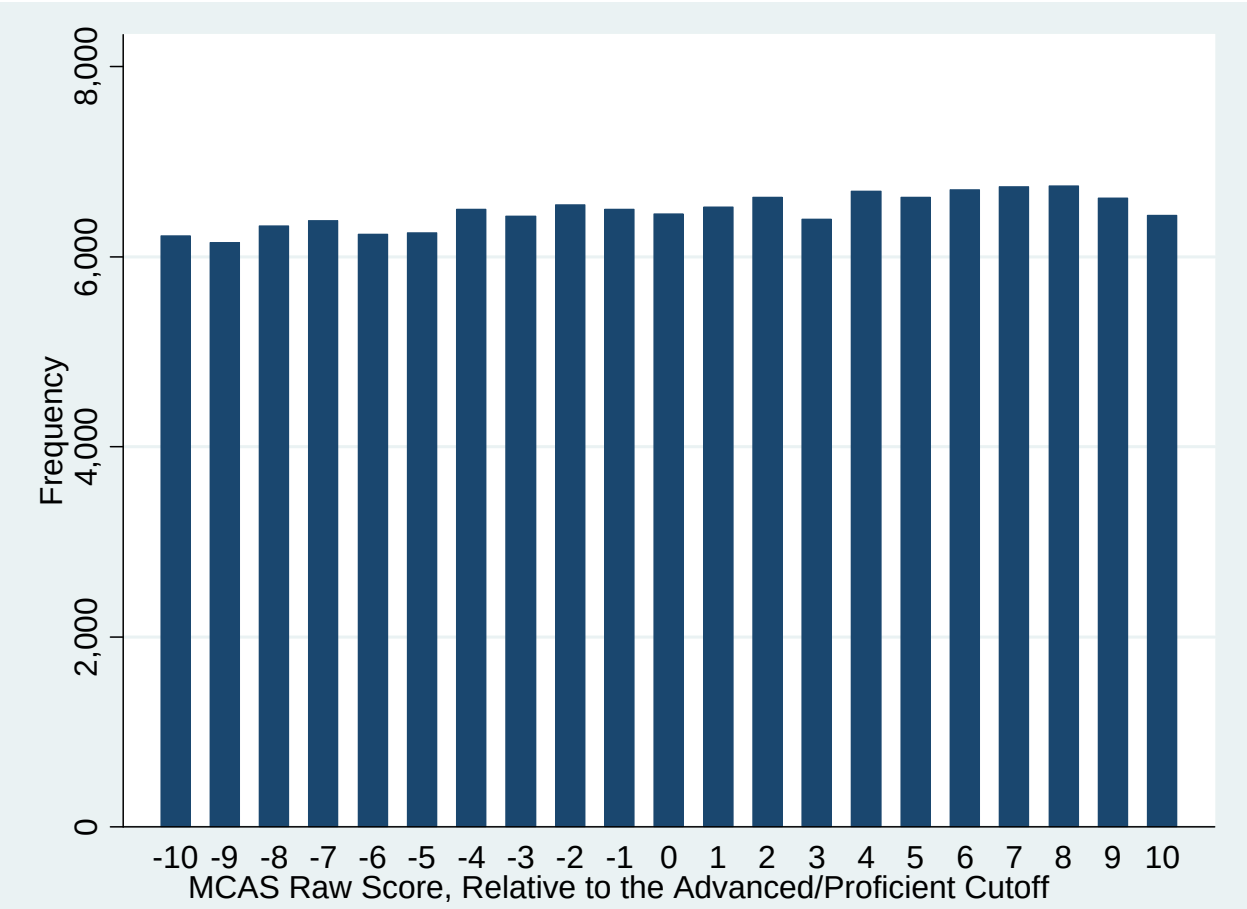

Fitted Probability of Attending College
By Four-Year College Aspirations
college
0
.2
.4
.6
.8
1
-10
-5
0
5
10
mathscore_adv
Four-Year College Aspirations
Not Four-Year College Aspirations

Figure 2. Density of student 10th grade mathematics MCAS scores, relative to the Advanced/Proficient cutoff.



Appendix A. Sample report provided to students with their MCAS test results.

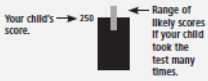
Your child's performance levels and scores

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:



Performance Level		English Language Arts	Mathematics	Science and Technology/ Engineering
Advanced Students at this level demonstrate a comprehensive and in-depth understanding of challenging subject matter and provide sophisticated solutions to complex problems.	280			
	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.