NBER WORKING PAPER SERIES

HOW PERFORMANCE INFORMATION AFFECTS HUMAN-CAPITAL INVESTMENT DECISIONS: THE IMPACT OF TEST-SCORE LABELS ON EDUCATIONAL OUTCOMES

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

Working Paper 17120
http://www.nber.org/papers/w17120

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2011

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, the Spencer Foundation, or the National Bureau of Economic Research. 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).

How Performance Information Affects Human-Capital Investment Decisions: The Impact of Test-Score Labels on Educational Outcomes
John P. Papay, Richard J. Murnane, and John B. Willett
NBER Working Paper No. 17120
June 2011
JEL No. I20,I21,J24

ABSTRACT

Students receive abundant information about their educational performance, but how this information affects future educational-investment decisions is not well understood. Increasingly common sources of information are state-mandated standardized tests. On these tests, students receive a score and a label that summarizes their performance. 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. Consistent with a Bayesian-updating model, these effects are concentrated among students with weaker priors, specifically those who report before taking the test that they do not plan to attend a four-year college.

John P. Papay
Brown University
Education Department
Providence, RI 02912
john_papay@mail.harvard.edu

Richard J. Murnane
Graduate School of Education
Harvard University
6 Appian Way - Gutman 469
Cambridge, MA 02138
and NBER
richard_murnane@harvard.edu

John B. Willett
Graduate School of Education
Harvard University
6 Appian Way - Gutman 412
Cambridge, MA 02138
John_Willett@harvard.edu

How Performance Information Affects Human-Capital Investment Decisions: The Impact of Test-Score Labels on Educational Outcomes

Standard economic models posit that students invest in further education if the expected marginal benefits of the investment exceed the expected marginal costs. Over time, as students accrue information about their educational performance, their perceptions of their abilities change. Because students' abilities determine both the benefits and the costs of further education, these evolving perceptions may influence decisions about future educational investments. A simple model of Bayesian updating fits this dynamic well, as in each period students have prior beliefs about the value of future investments but update these beliefs as they obtain additional information about their performance.[1] Evidence suggests that students indeed use performance data to update their plans about continuing in school (Jacob & Wilder, forthcoming).

Educational investment decisions are important because educational attainments are strong predictors of subsequent labor-market earnings (Goldin & Katz, 2008). However, for individuals, these decisions can be complicated matters that involve processing—explicitly or implicitly—a great deal of information. 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 reform in American public education has increased dramatically the amount of available information, particularly about students' mathematics and reading skills.

Recently, economists have paid a great deal of attention to the processes by which individuals make decisions when faced with an abundance of information. Theories of bounded rationality suggest that the cognitive (or time) cost of processing large amounts of information

[1] Similarly, students may gain additional information about the expected returns to schooling or the cost of additional schooling; in this paper, we focus on information about students' abilities and performances.

may exceed the benefit (e.g., Simon, 1957; Conlisk, 1996), leading individuals to use what Gigerenzer & Selten (2001) call “fast and frugal heuristics” in making decisions. Often times, these cognitive shortcuts enable people to make sufficiently good decisions by using only a fraction of the information available to them.

In this paper, we examine how information that students receive about their academic performance affects their decisions to enroll in post-secondary education. In particular, we look at one specific piece of data – student performance on the state standardized mathematics test in grades 8 and 10 in Massachusetts. One key feature of such test-based accountability systems is that every student receives not only a test score but also a label based on their performance (for example, *Failing*, *Needs Improvement*, *Proficient*, or *Advanced*). The state assigns the labels by determining three cut-points with which it divides the finer-grained test-score distribution into four regions. Given that understanding detailed test information can be a costly task, it makes sense to have a parsimonious summary that is easy for parents and students to interpret.

One feature of these performance labels is that students who are essentially equally skilled, but whose scores on the examination fall just on opposite sides of a cut point, receive different labels. This would not matter if students made use of all available information in assessing their skills, and if their parents and teachers did so as well, because the label provides no information beyond the fine-grained score. However, because the label provides a powerful summary of student performance—perhaps one layered with substantial emotional content—students (and their parents and teachers) may well respond to *it* rather than to the underlying test score. In this paper, we ask whether the label itself causes students to alter their decisions about pursuing post-secondary education. Importantly, we examine labeling on tests that carry no official, state-defined consequences for students.

Most research examining the role of information in students' educational decision-making processes has relied on laboratory experiments or descriptive analyses. Here, we are able to draw causal conclusions about students' responses based on the natural experiments that are created implicitly by the performance-labeling system that the state employs. We use a regression-discontinuity design to examine whether being assigned a more positive label affects future outcomes for students on the margin. Building on our past work, we focus on urban, low-income students' performances on the state mathematics examination.[2] We find that students do respond to these test-score labels and update their decisions about post-secondary education based on new information about their abilities. For example, being labeled *Advanced* rather than *Proficient* on the 10th grade test increases by five percentage points the probability that urban, low-income students at the margin will attend college. Interestingly, these responses are much greater for students on the margin of earning extreme labels (i.e., *Warning* or *Advanced*) than for more moderate labels (i.e., *Needs Improvement* or *Proficient*).

In a Bayesian-updating model, such responses should be strongest among students who have weak prior beliefs about their abilities. We look for these heterogeneous responses by focusing on students who report before they take the test that they do not plan to attend a four-year college. Indeed, we find much greater responses for students in this group. For example, being classified as *Advanced* rather than *Proficient* on the 10th grade mathematics test increases by ten percentage points the probability that they will enroll in college.

These results suggest that updating does occur and that seemingly small interventions like test-score performance labels can have substantial effects, particularly for some students. As theory would predict, these effects are concentrated among students with weaker prior beliefs

[2] In an analysis of students' behavioral responses to failing a high-stakes exit examination, we found that effects were concentrated among urban, low-income students on the mathematics examination (Papay, Murnane, & Willett, 2010).

about their academic abilities. These results provide strong evidence that students (or their teachers) are not using the full range of performance information available to them but that the labels themselves induce important behavioral responses. Using these labels as summaries may work well for some students, but it has important consequences for students near the cut-offs. This paper raises important implications 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. 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.

Background and Context

Responses to abundant information

In many contexts, individuals have access to abundant information that can inform decision-making. Many traditional economic models assume agents take into account all available information in making their decisions. However, influenced by the work of Herbert Simon (1957), economists have introduced models of bounded rationality that include the costs associated with information processing. In these models, agents may rely on heuristics – rules of thumb or cognitive short-cuts – to make decisions (Conlisk, 1996). Economists and psychologists have also examined the role of emotional responses and self-image in economic decisions (e.g., Kaufman, 1999; Loewenstein, 2000; Ackerlof & Kranton, 2002). Muramatsu & Hanoch (2005) argue that emotions "play a central role in guiding and regulating choice behavior, by virtue of their capacity to modulate numerous cognitive and physiological activities" (p. 202). Emotional responses can affect how individuals use information and thus can influence decision-making processes. This may be particularly true for adolescents, who 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 emotions are changing rapidly (Nelson & Sheridan, forthcoming).

Educational performance and updating

Many economic models posit that the benefits and costs of investing in additional education depend on an individual's abilities.[3] Individuals who believe they have stronger academic abilities may invest more heavily in education. From a Bayesian perspective, the extent to which an individual updates perceptions of his abilities when presented with new information depends on the strength of his prior beliefs, the importance of the new information, and the extent to which the new information is consistent with prior beliefs.[4]

[3] For example, in both human capital and signaling models, student ability is an important factor in educational investment decisions.

[4] Performance labels might also affect students' decisions about education by influencing behaviors of parents and teachers. 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 affect student outcomes (e.g., Jussim & Harber, 2005). In fact, President 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. Furthermore, 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 operate in different directions, and we cannot disentangle them here.

A number of scholars have hypothesized that adolescents' beliefs about their academic abilities may be particularly susceptible to new information, and that perceptions of ability therefore must influence adolescents' educational decisions. 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 (e.g., Crocker et al., 2003; Shen & Pedulla, 2000). Claude Steele and Joshua Aronson's work on stereotype threat suggests that external factors can affect individuals' performances on cognitive tasks. 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).

These psychological studies largely rely on experimental manipulations in a laboratory setting. However, several recent observational studies provide descriptive evidence that students do update their expectations about how much education they will complete when they receive new information about their academic performances (Jacob & Wilder, forthcoming; Stinebrickner & Stinebrickner, 2009).[5] This may provide part of the explanation why a substantial percentage of students who report that they plan to attend college fail to do so (Jacob & Wilder, forthcoming).

We might expect substantial heterogeneity across students in how malleable their judgments of their abilities are (i.e., in the strength of their priors). As Rabin and Schrag (1999)

[5] 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 and use "expectations" or "plans" throughout the paper.

assert, “if we posit a cost to information processing, in many settings the natural stopping rule would be to process information until beliefs are sufficiently strong in one direction or another, and then stop” (p. 41). Thus, students with strong priors about their abilities may be quite insensitive to new information, while students whose beliefs are less fully developed may be more sensitive. In other words, students whose self-identities and beliefs about their abilities are strongly formed may not change these beliefs except in the face of abundant evidence to the contrary. In contrast, students who lack confidence might be much more willing to update their beliefs.

The amount of updating also depends on how important students perceive the new information to be. For example, an “A” on a small homework assignment may not influence a student’s ideas about her abilities as much as an “A” on a final course examination. Given that teachers, school administrators, and even state officials emphasize the importance of scores on state standardized tests (often because the scores have consequences for educators), students may see them as meaningful measures of their proficiency.

Understanding the role of performance data in influencing students’ decisions about whether to invest in post-secondary education is important for two related reasons. First, the college wage premium is substantial and has grown over the past few decades (Goldin & Katz, 2008). Consequently, it makes sense for public policies to support college enrollment for all students with the capacity to succeed in post-secondary education. Second, 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.

Standards-based reform and state testing

Over the past few decades, most states, including Massachusetts, have implemented test-based accountability programs to monitor student progress toward mastering content standards. 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.[6] Currently, states must test all students in both mathematics and English language arts (ELA) in grades 3 through 8, and on a single occasion in high school.

The NCLB legislation set the goal that all American public-school students should be proficient in mathematics and 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.

Although the testing policies under NCLB did not mandate any consequences for students who failed to demonstrate mastery of state content standards, some states have implemented "high-stakes" testing for students, particularly in high school. Currently, 26 states, serving nearly three-quarters of the nation's children, have or are phasing in exit examinations, typically in ELA and mathematics, that high school students must pass in order to graduate (Center on Education Policy, 2008). In earlier work, we and others have examined the effects of barely failing these exit examinations (Papay, Murnane, & Willett, 2010). Students who fail an

[6] 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."

exit examination must retake and pass it in order to graduate from high school. As a result, failure has tangible consequences for students.

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 college-planning and college-enrollment decisions 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 our analysis on low-income students attending urban schools. We do so because, in earlier work, we found that failing one of the 10th grade exit examinations reduced the probability of high school graduation for low-income urban students, but it did not do so for their suburban or wealthier peers (Papay, Murnane, & Willett, 2010). In Massachusetts, wealthier students and suburban students have several advantages over urban, low-income students, on average. Their families often have better access to outside academic supports. Furthermore, their schools, which tend to serve lower concentrations of students with limited English proficiency, special educational needs, and from impoverished backgrounds, are able to concentrate more resources on individual struggling students than urban schools with many such students can. In this paper, we examine the effects of labeling on very different tests – those that do not carry official consequences for students – but we retain our focus on this traditionally disadvantaged group. Similarly, we focus on testing in mathematics because our past work found that, for urban, low-income students on the margin of passing, barely passing the mathematics high-school exit examination increased the probability of graduation substantially, but there were no effects of barely passing the ELA examination (Papay, Murnane, & Willett, 2010).

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 educational plans that were asked before the students were administered 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, & Duncan, 1972; Sewell, Haller, & Ohlendorf, 1970; Sewell, Haller, & Portes, 1969). However, the development of educational expectations is a process that researchers do not understand well (Jacob & Wilder, forthcoming). We examine whether the effects of labeling depend on students' initial post-secondary educational plans, and whether the performance label students receive causes 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 unambiguously with our regression-discontinuity strategy. However, we use students' test-score histories to provide some 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 information that urban, low-income students receive on the Massachusetts state mathematics test affect their college enrollment decisions?

2: Are the post-secondary plans and college enrollment decisions of students who did not plan initially to attend a four-year college more sensitive to new performance information than the decisions of students with college-going plans?

3: Does prior test performance shed light on the relative importance of encouragement and discouragement 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) examinations in 1998. For most students, performance on these tests carries no official consequences. However, starting with the class of 2003, the 10th grade tests became high-stakes exit examinations that students must pass in order to graduate from high school.[7]

To address our research questions, we have integrated 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, and responses to surveys that students complete just before taking the MCAS examinations. We have supplemented this dataset with records from the National

[7] 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.

Student Clearinghouse that tracks 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 8th grade, the examination is used simply to hold schools and districts accountable. However, the 10th grade examination is a high-stakes exit examination. As a result, in 10th 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* vs. *Warning* on the 8th grade test, at *Proficient* vs. *Needs Improvement* on the 8th and 10th grade tests, and at *Advanced* vs. *Proficient* on the 8th and 10th grade tests. The state does not treat students on either side of these cutoffs differently and this information is not provided to colleges.[8]

We pool data across several years, examining students who took the 8th or 10th grade mathematics examinations in the spring of 2003 through 2007. These students are members of the graduating cohorts of 2005 through 2011. 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.

Measures

To address our research questions, we created several outcome variables. Our main outcome ($COLL_i$) measures whether students attended college by June 1, within one year after their intended cohort's high-school graduation date.[9] We created an additional outcome by

[8] Importantly, for some students, earning *Advanced* instead of *Proficient* 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. As we explain below, we focus our attention on a sample of students for whom this is not the case.

[9] 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 8th grade examination and two years after the student took the 10th grade examination.

recoding responses about their post-secondary educational plans that students provide immediately before they take the MCAS examinations. The survey question to which they responded reads as follows:[10]

> 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. 76% percent of Massachusetts low-income urban 10th grade students (and 85% of 8th 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.[11]

Our key predictors come from the state-testing dataset, which includes 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: as item-level responses, raw scores, scaled scores, and performance levels. The scaled scores 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:

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

[11] 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, for most students we cannot track their progress 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.

(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 mapping on to a single scaled score, we use raw scores in our analyses.[12]

Students receive information about their test performance in detailed reports several months after taking the test. In Appendix A, we include an example of one such report; notice that it provides students with information about their test score, a confidence interval around the score, and a performance label. It also contains interpretive information (not shown) to help students and parents make sense of their test scores. Thus, students and parents 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.

To implement our regression-discontinuity approach, we center students' raw scores by subtracting out the value of the corresponding minimum passing score associated with the relevant cut-point. On the re-centered continuous predictor ($MATH_i$), which serves as our "forcing variable" in the regression-discontinuity analyses, a student with a score of zero had achieved the minimum passing score at that cut-point. We also created a dichotomous version of this same predictor ($ABOVE_i=1\{MATH_i \geq 0\}$).[13] To address our third 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 (e.g., grade 8 for the grade

[12] 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 non-linear 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 (MA DOE, 2002, 2005).

[13] 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.

10 analyses).

Sample

As stated above, we focus our analyses 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.[14] This group constitutes 13% percent of Massachusetts 10th grade students who take the examination (and 18% of 8th graders). 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 8th 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 10th grade. 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 beginning with the 2005-06 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 second research question) and we cannot examine actual college-going outcomes for this group.

INSERT TABLE 1 ABOUT HERE

In several of our analyses, we examine heterogeneity based on the post-secondary educational plans that students report before they take the 8th and 10th grade examinations. Among urban, low-income students, 63% of those who completed the 10th grade survey, and 57% of those who completed the 8th grade survey, reported that they planned to attend a four-year college. 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.

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

Note that these figures do not correspond to response rates for 8th graders because the survey questions were not asked every year.

INSERT TABLE 2 ABOUT HERE

In the third column, we present the sample percentage 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% of low-income urban 10th graders who score *Advanced* plan to attend a four-year college, compared to just 46% 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 10th 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, so these results do not simply reflect the effect of differential performance across demographic groups (results available from authors on request).

INSERT FIGURE 1 ABOUT HERE

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 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:

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

If the cut score were established exogenously, then students just on either side of the cut score must be equal in expectation prior to labeling and the estimated difference between these parameters provides an unbiased estimate of the causal impact of the classification for students at the cut score, in the population (Lee & Lemieux, 2010; Murnane & Willett, 2011; Shadish, Cook, & Campbell, 2004). Because the labels are applied rigidly such that all students who score below the cut-off on the forcing variable are assigned one label and all students who score above the cut-off 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 post-secondary plan outcome. In its basic formulation, this approach involves fitting a linear probability model of the following form:

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

for the i^{th} 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 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 two key assumptions. First, students must not be able to manipulate their position on the forcing variable knowingly relative to the cut score. Given the complicated scaling procedures used to determine the cut-offs, we have strong reason to believe that this assumption holds. Second, we 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 on to 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, & Van der Klaauw (2001).[15]

Our implementation of this strategy follows the approach laid out by Imbens and Lemieux (2008). 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 & Lemieux (2008).[16] 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

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

[16] $h* = \arg \min_{h} \frac{1}{N}\sum_{i=1}^{N}(\hat{COLL}_i(h) - COLL_i)^2$, 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.

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.[17]

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 (1) a vector of covariates describing selected aspects of the student background, including dichotomous predictors that describe student race, gender, and whether the student was new to the state, was currently or formerly classified as limited English proficient, or required special education. 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 second research question, we fit a statistical model similar to that specified in (1), using a similar local linear-regression approach. In this case, though, we include the student's pre-treatment self-reported college plans ($COLL\_PLAN_i$) as a covariate and interact it with the main predictors, as follows:

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

for the i^{th} 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

[17] 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.

[18] We tested whether adding school fixed effects would increase the explanatory power of our models, and found that they 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. Results are available from the authors on request.

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 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 third 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 grade 8 and grade 10. We find relatively 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. Again, we focus all of our analyses on tests that carry no official consequences for Massachusetts students.[19]

Findings

(1) Does the performance label affect post-secondary enrollment decisions?

[19] The state has a college scholarship program (called the *Adams Scholarship*) that provides post-secondary support for students with high 10th grade MCAS 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% of all test-takers in their district in terms of total score. To avoid any potential confounding of the effects of scholarship receipt and performance labeling, we focus our attention in 10th grade on the sample of students for whom scoring *Advanced* instead of *Proficient* in mathematics does not affect their scholarship eligibility. We do this in two ways. First, we 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% of their district. Both samples give us quite similar results, so we present our findings using the less restrictive, second sample. Results from the first sample are available from the authors on request.

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 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 8th grade increases the fitted probability of enrolling in college by 2.1 percentage points (*p*=0.056). Since only 38% percent of urban, low income 8th 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* in either grade. Interestingly, this is the cutoff that is used to define Adequate Yearly Progress under *No Child Left Behind*. Students near this cut score may be subject to strategic behaviors, and the focus on these "bubble kids" on either side of the cutoff may produce these dampened effects (Booher-Jennings, 2005; Neal & Schanzenbach, 2010). Alternately, more moderate labels like *Proficient* and *Needs Improvement* may not be as meaningful to students.

We find that receiving the *Advanced* rather than the *Proficient* label on the 10th 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% of the urban, low-income students scoring near the cutoff enroll in college. In contrast, receiving *Advanced* rather than the *Proficient* label on the 8th grade mathematics test has no impact on the probability of college enrollment. One possible explanation for the difference between the 8th and 10th grade results concerns the test itself. Only 3.6% of Massachusetts low-income urban students earn an *Advanced* rating in 8th grade, compared to 13% of 10th graders. Consequently, students scoring near the *Advanced/Proficient* cut-score in 8th grade are quite high-performing.

INSERT TABLE 3 ABOUT HERE

(2) 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. If students respond to the information embedded in performance labels, we expect these responses to be concentrated among students who have weaker prior beliefs about their academic abilities. 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 college-going plans. For students who did not plan to attend a four-year college, earning a more positive label has a substantial, positive effect on their college enrollment decisions.

Importantly, our analysis of the 8th grade results is limited by data availability. The state first administered the 8th 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 10th grade. The results are striking. As seen in Table 4 being classified as *Needs Improvement* instead of *Warning/Failing* on the 8th grade mathematics examination raises students' fitted probability of expressing four-year college as their intended post-secondary goal on the 10th grade survey by 4 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 traditional levels of statistical significance (*p*=0.238). Finally, scoring *Advanced* instead of *Proficient* in 8th 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 8th 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 10th 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.

INSERT TABLE 4 ABOUT HERE

We find very similar patterns on the 10th 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 *Proficient*, increases the fitted probability that they will attend college by 9.9 percentage points (p=0.010). For these 10th graders on the margin, the estimated probability of attending college increases from approximately 39% to 49% simply by being labeled *Advanced* instead of *Proficient*.

In Figure 2, we present these results visually. In each figure, we include the sample probabilities of attending college (grade 10 outcome) or expressing four-year college-going plans (grade 8 outcome) 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.

INSERT FIGURE 2 ABOUT HERE

These figures illustrate several important patterns. First, students who report as 10th graders that they plan to attend a four-year college do indeed attend college at a substantially higher rate than students with other plans. Similarly, students who report as 8th graders that they plan to attend a four-year college have a much greater probability of reporting the same plans in 10th grade. Second, across all three cut-scores, there is no clear disruption in the smoothed relationship for students with four-year college expectations. 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 8th and 10th grades. By contrast, 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' post-secondary decisions than labeling at the top or bottom.

(3) Does prior test performance shed light on the relative importance of encouragement and discouragement effects?

The findings presented above indicate clearly that the information embedded in performance labels causes students to update their ideas about their educational futures and to alter college-going decisions. Students without plans to attend a four-year college are most liable to update their plans and alter their decisions. However, the precise interpretation of these findings proves challenging because 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 update positively their beliefs about their abilities, thereby increasing the probability they subsequently attend college. However, relatively high-performing students who are labeled as simply *Proficient* may be discouraged by their failure to achieve the more prestigious *Advanced* label and, as a result, may not consider themselves "college material". This would represent a negative updating of their abilities. 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. Here, we assume that students respond to the information embedded in the test performance label when it is different from the label that they had earned in a previous grade. However, we assume that no updating occurs if the new information matches students' prior labels. We present results from these analyses for the 8th 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 8th 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 10th grade test, the performance level cutoffs are lower than on the 8th grade examination. For example, nearly all students near the *Advanced/Proficient* had scored *Proficient* or lower on the 8th grade test. Thus, in both 8th and 10th grades the encouragement effect appears to predominate at the top of the distribution.

INSERT TABLE 5 ABOUT HERE

Threats to Validity

As we have noted, the internal validity of a regression-discontinuity design depends on two important assumptions. First, the "treatment" – here embodied in the particular performance label applied to students scoring immediately at/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. In other words, students must not be able to manipulate their position on the forcing variable relative to the cut score. If this condition holds, then all student characteristics, both observed and unobserved, should be a smooth function of the forcing variable around the cut score.

In our research, we argue that this assumption holds 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 hard, if not impossible, for students at the margin of passing to manipulate their explicit positions with respect to the cut-off knowingly while taking the examination. However, we can also test if this assumption is met in several ways. First, if students can influence their position on the forcing variable relative to the cut-off by manipulating their test performance, we would expect the test-score distribution to be discontinuous near the cut-off. In all cases, we see no discontinuity apparent at the cut score.[20] Second, if students can influence their labeling after taking the test, we would expect to see some non-compliers, or 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 extensive exploratory analyses to check for smoothness in the relationship between observed student characteristics and the forcing variable around the cut-off. Visual inspection of the underlying distributions suggests that these relationships are smooth over the cut-score. Seeking to summarize these analyses in a single test, we followed the approach suggested by Lee & Lemieux (2010). We fit a set of seemingly unrelated regression (SURE) 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 cannot reject the null hypothesis and find no reason to doubt that the state has imposed the cut score exogenously.

INSERT TABLE 6 ABOUT HERE

The second 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 an optimal bandwidth of 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.

[20] Figures available from the authors on request.

Therefore, 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 cut-off 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 3, we explore this result in more depth for both the 8th grade (top panel) and 10th grade (bottom panel) tests. Here, we present the estimated causal effect from our local linear-regression analysis with bandwidths ranging from 3 to 15 score points, along with 90% 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 8th grade, some of these estimates do not reach traditional levels of statistical significance, but the general pattern remains.

INSERT TABLE 7 AND FIGURE 3 ABOUT HERE

Discussion

We conclude that urban low-income students – or those who teach or advise them – do indeed respond to performance labels. In other words, the labels are used – or misused – as a "fast and frugal heuristic" that influences decisions even though they provide no additional information beyond that in fine-grained scores. 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. These labels matter even though students also receive the fine-grained test scores on which they are based. The effect of labeling is concentrated at the extremes of the test-score distribution. In other words, being labeled as *Warning* or as *Advanced* seems to matter a great deal to students; by contrast, being labeled as *Proficient* instead of *Needs Improvement* matters less. Furthermore, as theory predicts, 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. Students are updating their college-going plans in response to performance information, and this updating is concentrated among students with the weakest priors. The large labor-market returns to post-secondary education make the results important substantively.

There are at least two complementary explanations for the powerful effects of labels. First, cognitive limitations may make interpretations of complicated test-score data difficult and may increase students' reliance on the performance labels. The state attempts to minimize this issue by presenting test-performance data in a variety of ways (see *Appendix A*), including a visual depiction with error bars on the interpretive material. But, students—or their parents—may not have the skills necessary to understand these distinctions clearly.

Second, 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. 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. If anything, the fact that so seemingly weak a signal as the performance label on a state test can have such persistent and substantial effects on educational outcomes speaks loudly to the vulnerability of students' conceptions of their own abilities. In other words, urban, low-income students' priors about their educational abilities appear to be rather weak, even in the 10th grade. In particular, high-achieving students seem to respond quite strongly to external acknowledgment of their intellectual abilities.

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. 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, using explicitly 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. 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.[21] 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.

Furthermore, the fact that labeling matters more for students at the top or the bottom of the test-score distribution suggests two complementary explanations. First, more extreme labels such as *Warning* or *Advanced* may simply matter more to students or to their teachers. Alternately, schools may be involved in strategic behaviors to get students over the proficiency

[21] Results from models with school fixed effects produce quite similar point estimates to those presented here, although the estimates are somewhat less precise. Specific results are available from the authors on request.

cutoff and to keep them there. To the extent that schools focus attention on "bubble kids" on either side of the *Proficient* cutoff, we would expect any effects of labeling to be muted.

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 cut-off. However, if individuals respond to performance labels on these same 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 fact 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. 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. The performance label– ostensibly a fairly weak signal – has a powerful effect on student outcomes, including college enrollment decisions that occur several years after the test. This is particularly true for the encouragement of earning an *Advanced* label for relatively high-performing students. Furthermore, the labels we examine provide information about student performance relatively late in their academic careers. That students respond to these labels suggests that their priors are still relatively weak and that high school is not too late for potentially effective educational interventions.

These effects are particularly large for students who do not plan to attend a four-year college after high school. This group appears to be vulnerable to the effects of labeling, suggesting that these students are sensitive either to positive encouragement or negative reinforcement of their attitudes. That the responses to labeling appear to be positive encouragement effects at the top of the distribution suggests that the need to address this consequence is not as urgent for high-performing students. However, at the bottom of the distribution, earning a worse label appears to discourage students, suggesting that policymakers and school officials should consider finding ways to support those students who earn the “Warning” label. 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. 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.

Akerlof, G.A. and Kranton, R.E. (2002). Identity and schooling: Some lessons for the economics of education. *Journal of Economic Literature, XL*(December), 1167-1201.

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.

Conlisk, J. (1996). Why bounded rationality? *Journal of Economic Literature, 34*(June), 669-700.

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.

Gigerenzer, G. & Selten, R. eds. (2001) *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Goldin, C. & Katz, L.F. (2008). *The Race between Education and Technology*. Cambridge, MA: Belknap Press.

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., & Wilder, T. (forthcoming). Educational expectations and attainment. Chapter in a forthcoming volume edited by Greg Duncan and Richard J. Murnane (Russell Sage).

Jussim, L. and Harber, K.D. (2005) Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131-155.

Kaufman, B.E. (1999). Emotional arousal as a source of bounded rationality. *Journal of Economic Behavior and Organization, 38*, 135-144.

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

Lee, D.S. & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, *48*(2), 281-355.

Loewenstein, G. (2001). Preferences, behavior, and welfare: Emotions in economic theory and economic behavior. *American Economic Review: AEA Papers and Proceedings, 90*(2), 426-432.

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.

Muramatsu, R. & Hanoch, Y. (2005). Emotions as a mechanism for boundedly rational agents: The fast and frugal way. *Journal of Economic Psychology, 26*, 201-221.

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

Neal, D. & Schanzenbach, D.W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics, 92*(2), 263-283.

Nelson, C.A. & Sheridan, M.A. (forthcoming). Lessons from neuroscience research for understanding causal links between family and neighborhood characteristics and educational outcomes. Chapter in a forthcoming volume edited by Greg Duncan and Richard J. Murnane (Russell Sage).

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

Rabin, M. & Schrag, J.L. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics, 114*(1), 37-82.

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.

Simon, H.A. (1957). *Models of Man: Social and Rational.* New York: John Wiley and Sons, Inc.

Stinebrickner, T.R. & Stinebrickner, R. (2009). Learning about academic ability and the college drop-out decision. *NBER Working Paper 14180.* Retrieved from August 17, 2010 from http://www.nber.org/papers/w14810.

Table 1. Description of data structure and cohorts, by the availability of specific outcomes.

Test Cohort	College-going plans	College attendance
8th Grade:		
2002-03	2004-05	2007-08
2003-04	2005-06	2008-09
2004-05	2006-07	--
2005-06	2007-08	--
2006-07	2008-09	--
10th Grade:		
2002-03	--	2005-06
2003-04	--	2006-07
2004-05	--	2007-08
2005-06	--	2008-09

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 8th grade (2002-03 to 2007-08) and 10th grade (2002-03 to 2006-07)

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**
8th Grade:			
Advanced	2,015	876	84.70
Proficient	8,068	3,358	74.42
Needs Improvement	19,744	7,328	62.83
Warning	39,402	11,895	46.73
10th Grade:			
Advanced	4,699	4,014	84.93
Proficient	7,073	5,721	70.98
Needs Improvement	11,854	9,337	60.03
Failing	12,573	8,486	45.60

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. Cell entries include the parameter estimate, standard error (in parentheses), optimal bandwidth used, sample size, and approximate *p*-value.

Outcome	***Needs Improvement/ Warning***	***Proficient/ Needs Improvement***	***Advanced/ Proficient***
Panel I: 8th Grade			
College attendance	0.021~	0.001	0.007
	(0.009)	(0.028)	(0.035)
	h=3	h=8	h=4
	5,801	6,313	1,248
Panel II: 10th Grade			
College attendance	N/A	0.008	0.051*
		(0.010)	(0.020)
		h=6	h=8
		8,280	4,171

NOTE: ~, $p<0.10$; *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. 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.

Table 4. 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 plans to attend a four-year college after high school. Cell entries include parameter estimates, standard errors (in parentheses), and approximate *p*-value.

Outcome	**Students with College Plans**	**Students without College Plans**	**Sample Size**
Panel I: 8th Grade			
Needs Improvement/Warning Cutoff:			
Express 4-year college plans (grade 10)	0.009	0.040~	3,824
	(0.031)	(0.021)	h=5
Proficient/Needs Improvement Cutoff:			
Express 4-year college plans (grade 10)	-0.028	0.062	4,487
	(0.012)	(0.051)	h=8
Advanced/Proficient Cutoff:			
Express 4-year college plans (grade 10)	0.007	0.137~	2,294
	(0.023)	(0.071)	h=8
Panel II: 10th 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

NOTE: ~, *p*<0.10; *, *p*<0.05; **, *p*<0.01; ***, *p*<0.001. 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.

Table 5. 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 scored above or below the cutoff on an earlier test. Cell entries include parameter estimates, standard errors (in parentheses), and approximate *p*-value.

Outcome	**Students with lower scores on prior test**	**Students with higher scores on prior test**	**Bandwidth**
Panel I: 8th 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	

NOTE: ˜, $p<0.10$; *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. 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.

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

NOTE: Covariates treated as outcomes include student race, gender, and whether the student was new to the state, was currently or formerly classified as limited English proficient, or required special education.

Table 7. Estimated effect of earning the more positive performance label at different cutoffs and on different outcomes, by bandwidth.

Outcome	**Bandwidth**				
	h*-2	**h*-1**	**h***	**h*+1**	**h*+2**
8th 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)
8th 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	0.04 ~	0.023	0.027
	(0.029)	(0.027)	(0.021)	(0.023)	(0.021)
8th 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)
8th 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)
8th Grade *Advanced/Proficient* cutoff – All Urban, Low-Income Students:					
College Attendance (h*=9)	--	-0.047	0.007	0	0.012
		(0.024)	(0.035)	(0.026)	(0.027)
8th 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)
10th 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)
10th 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)
10th 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)
10th 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)

NOTE: *, *p*<0.05; **, *p*<0.01; ***, *p*<0.001. 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 1. Sample probabilities of expressing an interest in attending a four-year college, by race and gender, for urban, low-income students in Massachusetts.

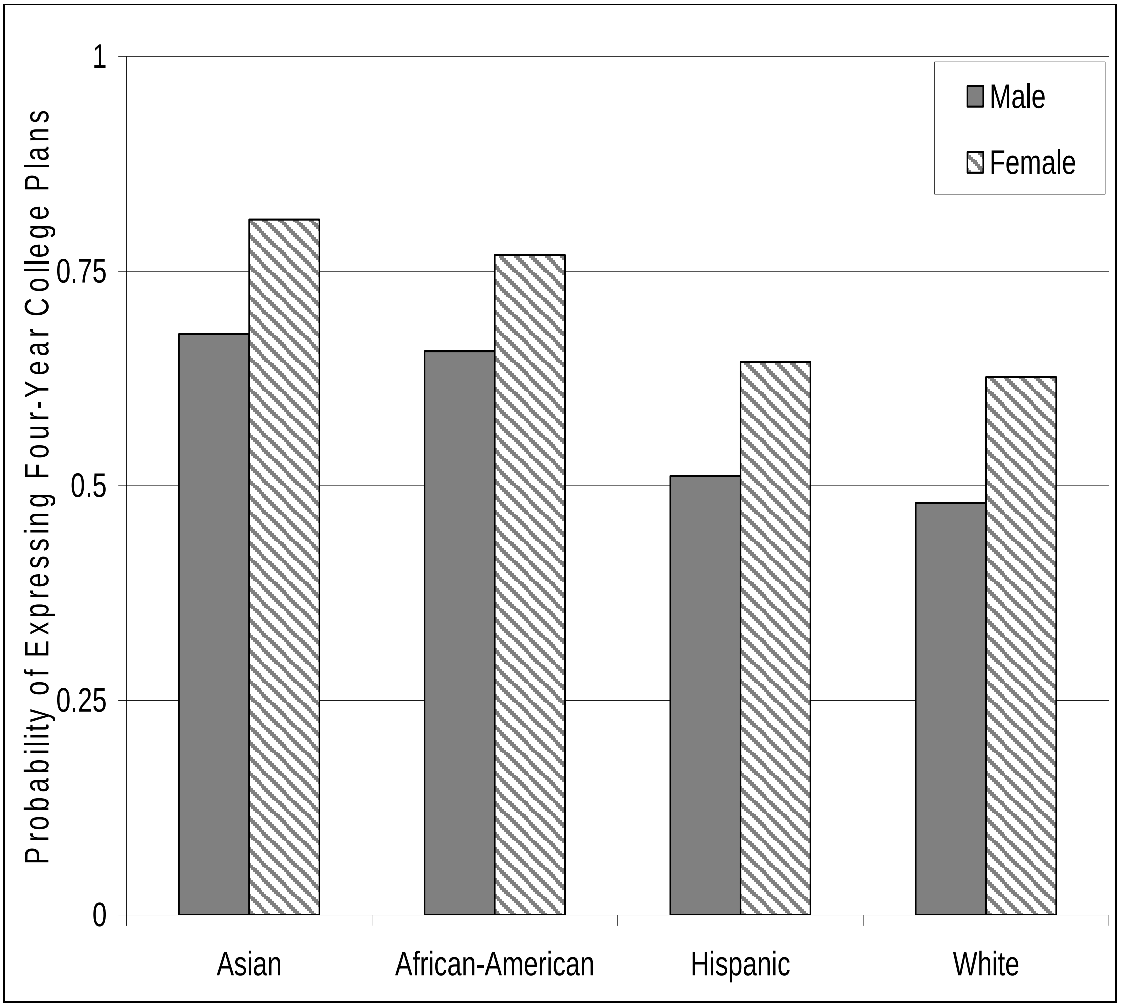
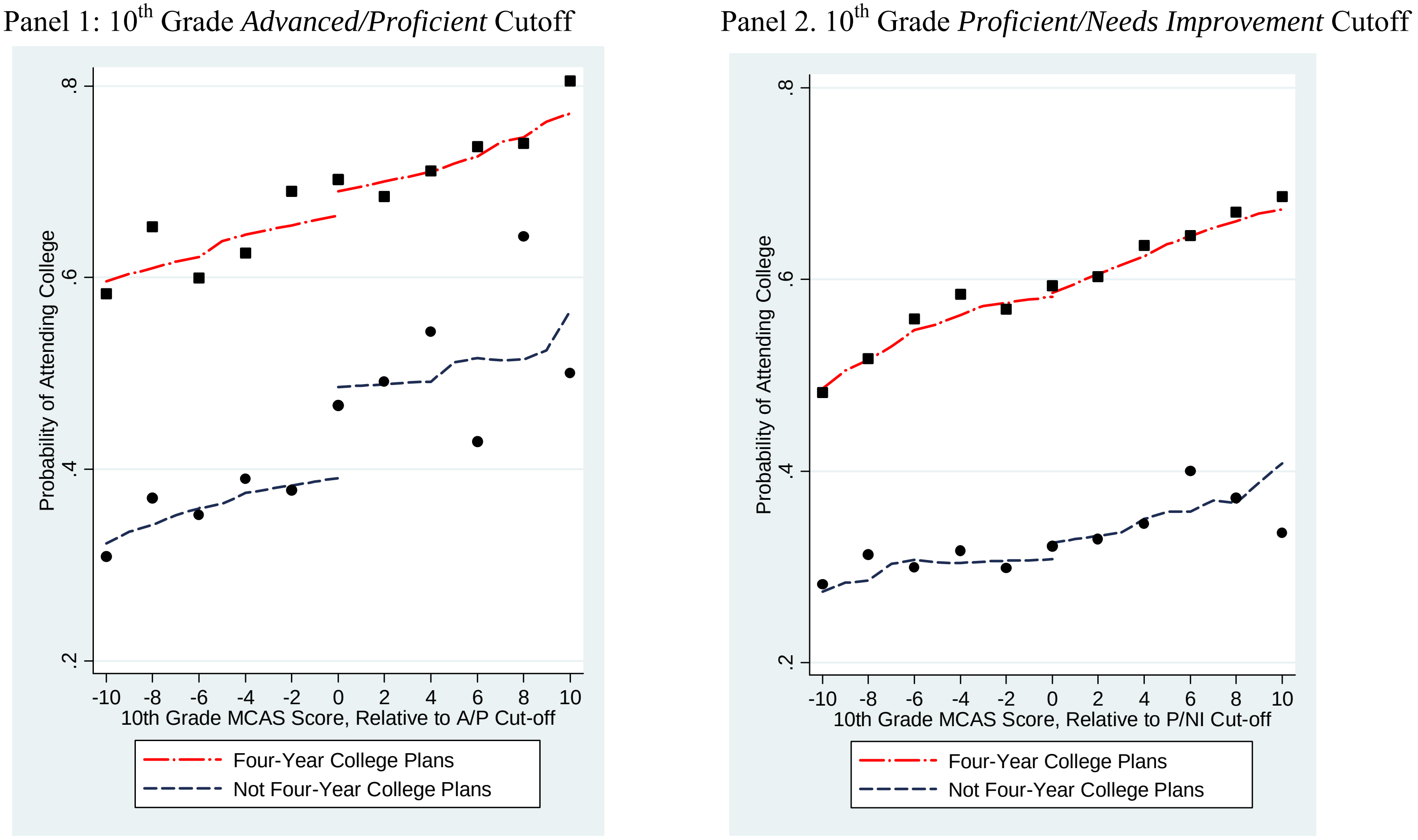
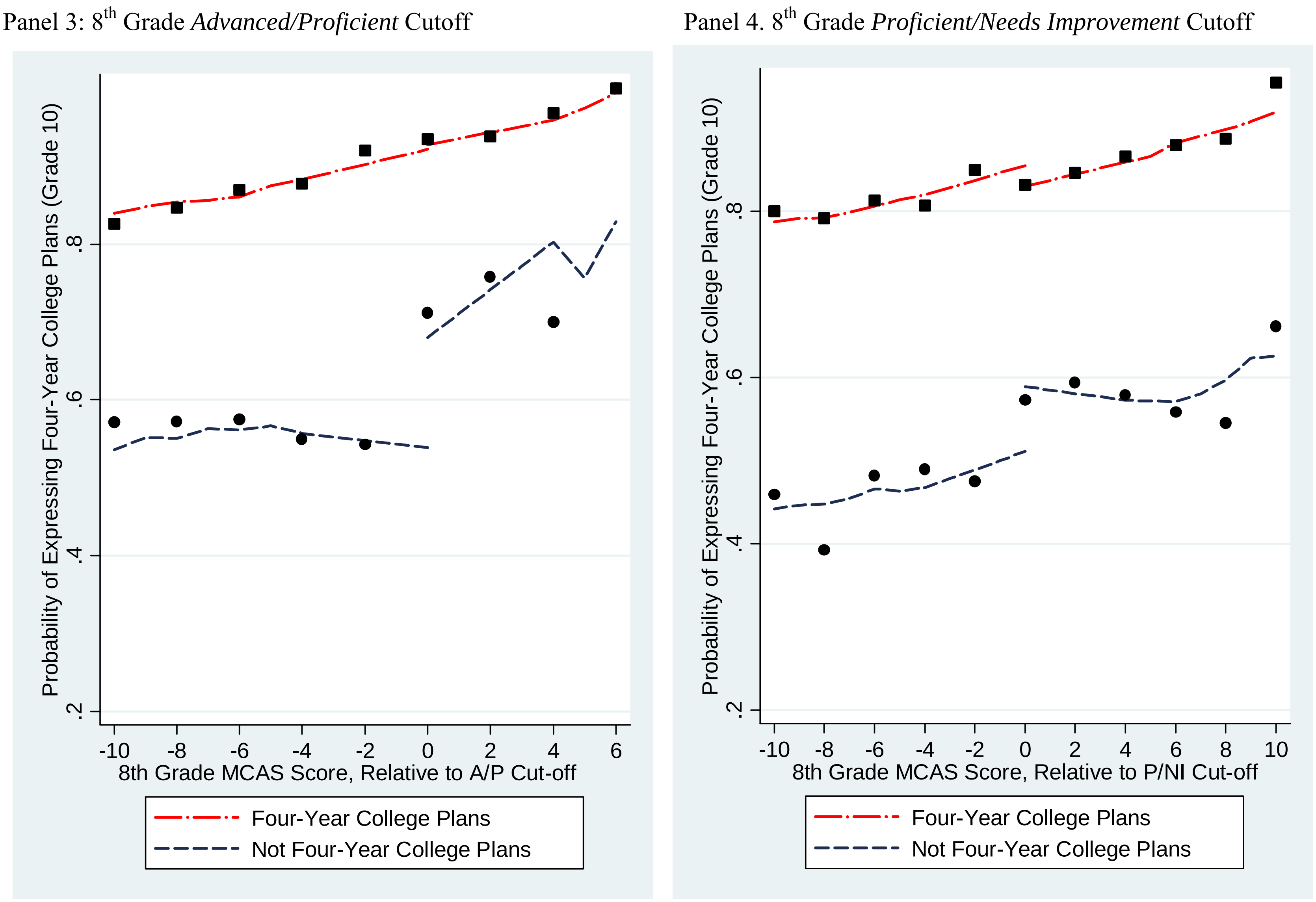


Figure 2. Fitted local linear-regression relationships between the probability of attending college and 10th grade mathematics score relative to the *Advanced/Proficient* cutoff (panel 1, *h**=8), the probability of attending college and 10th grade mathematics score relative to the *Proficient/Needs Improvement* cutoff (panel 2, *h**=8),the probability of expressing four-year college-going plans in grade 10 and 8th grade mathematics score relative to the *Advanced/Proficient* cutoff (panel 3, *h**=8), the probability of expressing four-year college-going plans in grade 10 and 8th grade mathematics score relative to the *Proficient/Needs Improvement* cutoff (panel 4, *h**=8), and the probability of expressing four-year college-going plans in grade 10 and 8th grade mathematics score relative to the *Needs Improvement/Warning* cutoff (panel 5, *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.



Panel 1: 10th Grade *Advanced/Proficient* Cutoff

Panel 2. 10th Grade *Proficient/Needs Improvement* Cutoff



Panel 3: 8th Grade *Advanced/Proficient* Cutoff

Panel 4. 8th Grade *Proficient/Needs Improvement* Cutoff

Panel 5: 8th Grade *Needs Improvement/Warning* Cutoff

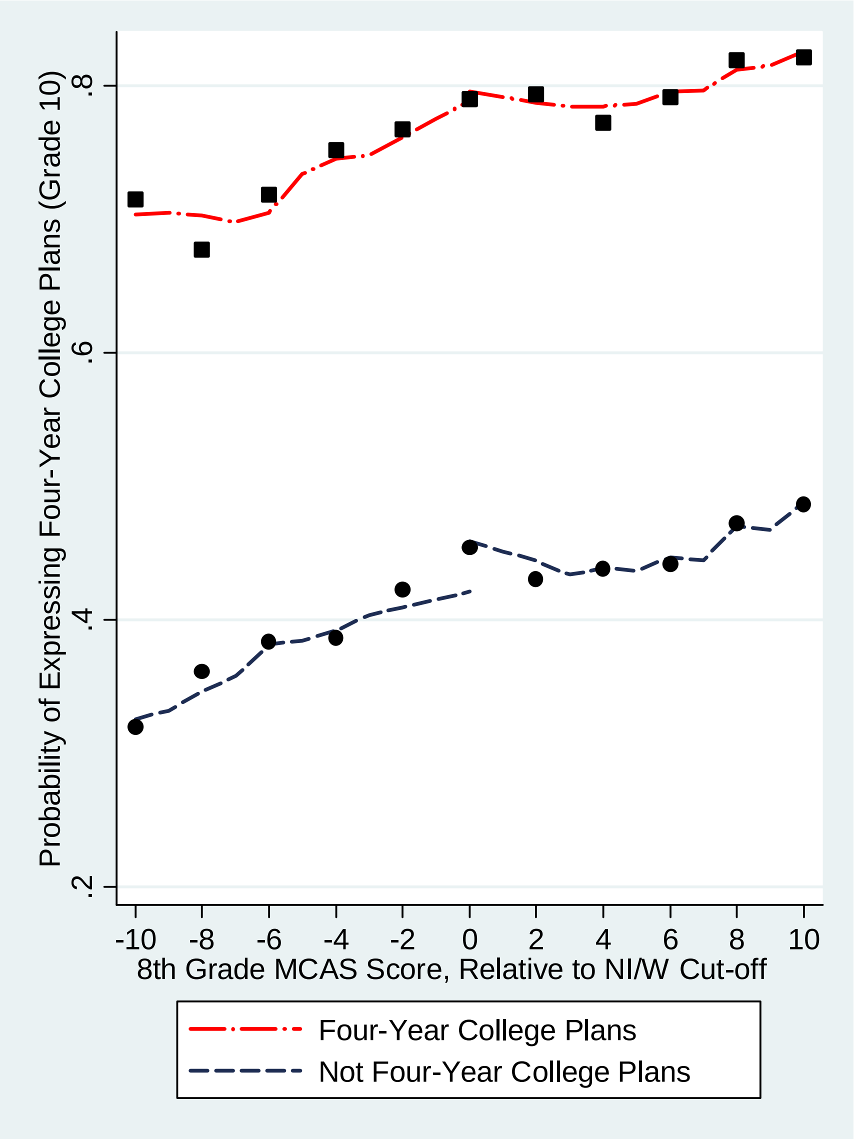
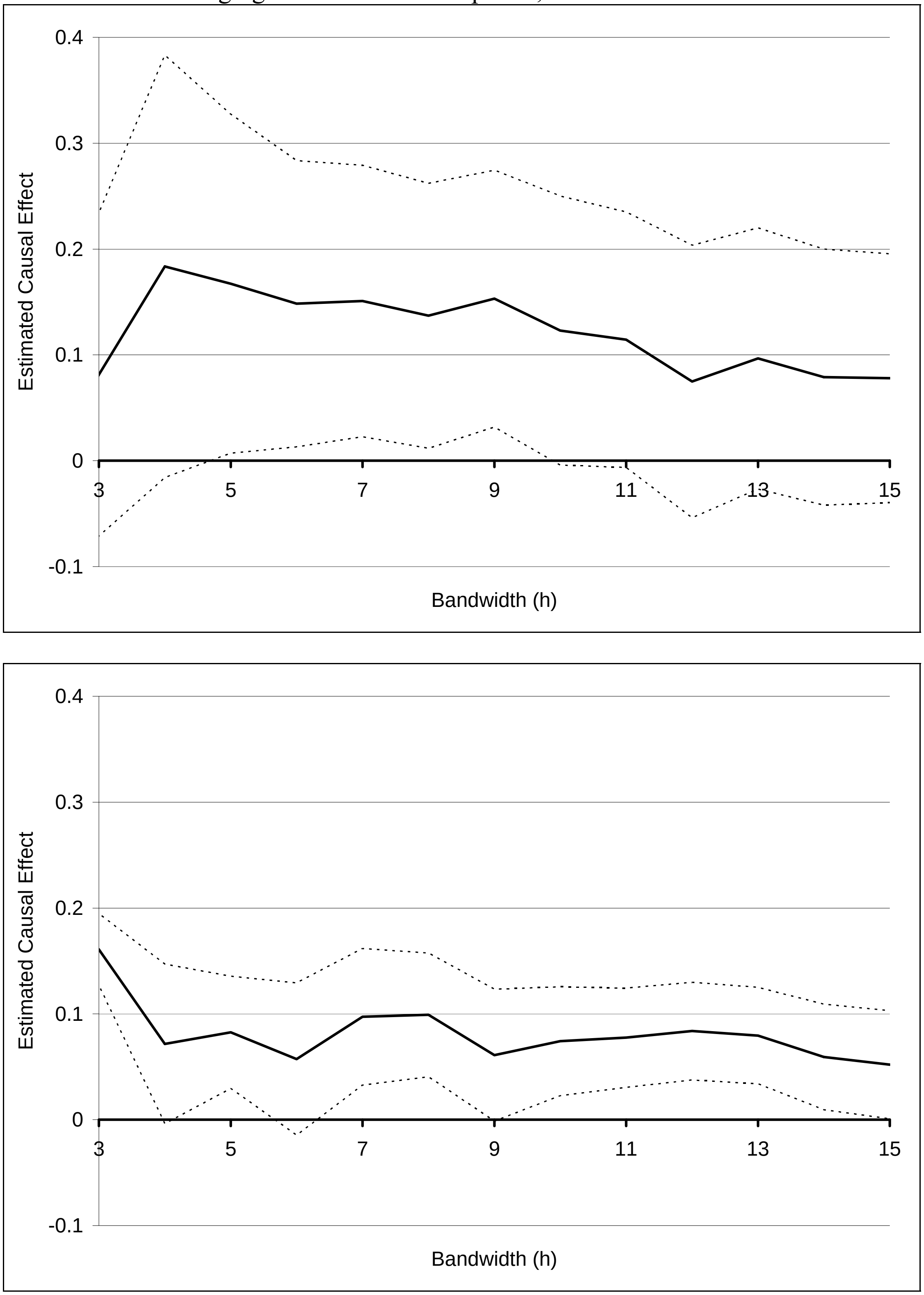
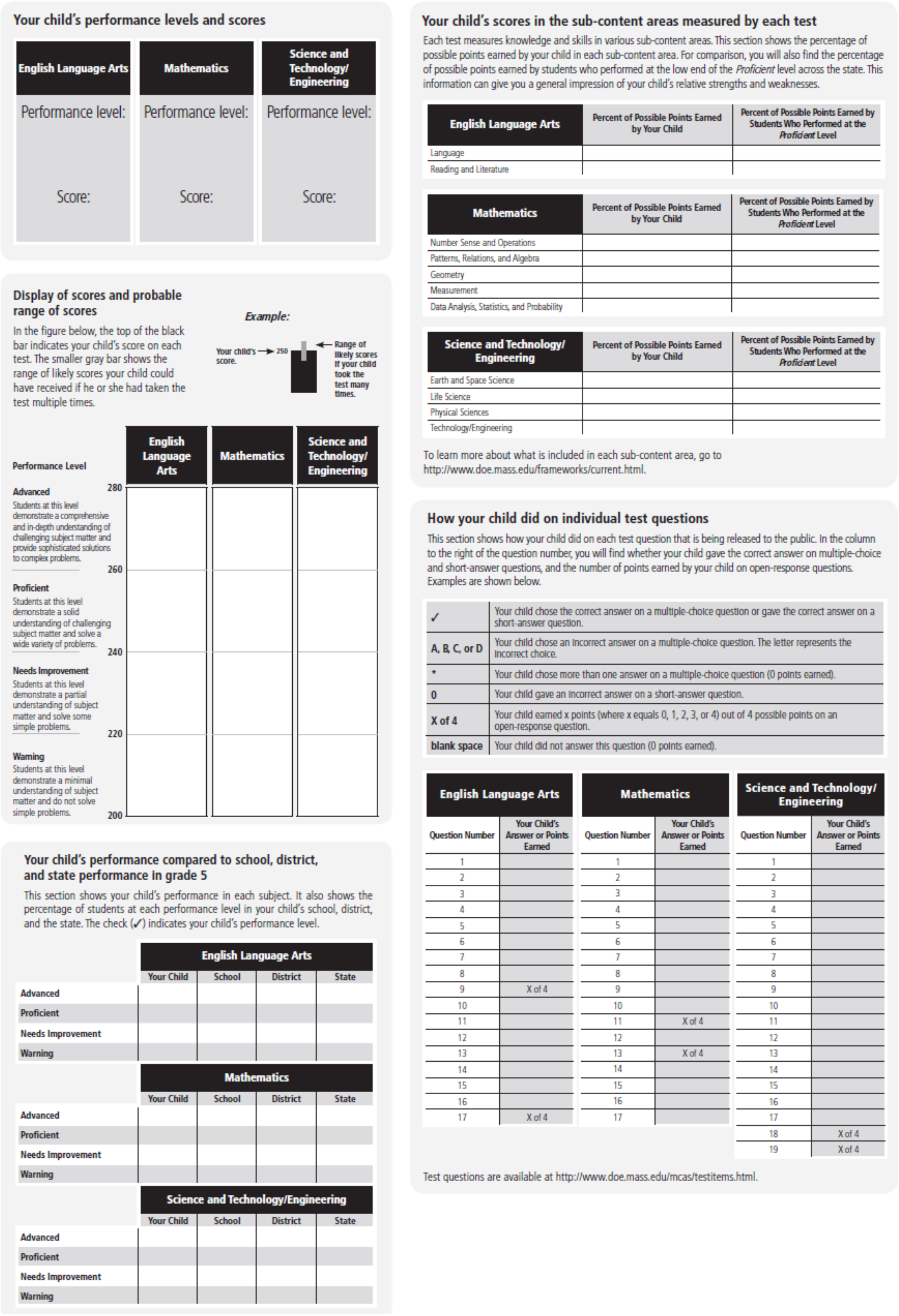


Figure 3. Estimated causal effect of earning *Advanced* instead of *Proficient* on the 8th grade (top panel) and 10th grade (bottom panel) tests, from local linear regression-discontinuity analysis with bandwidths ranging from 3 to 15 score points, with 90% confidence intervals.



*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:
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