Abstract Title Page
Not included in page count.

Title: The High-Stakes Effects of "Low-Stakes" Testing

Author(s): John P. Papay, Richard J. Murnane, John B. Willett

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Standard economic models posit that students invest in further education if the expected marginal benefits exceed the expected marginal costs. Over time, as students accrue information about their educational performance, their perceptions of their abilities evolve. Because students' abilities help determine the benefits and 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. Evidence suggests that students indeed use performance data to update their plans about continuing in school (Jacob & Wilder, 2010).

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.

Economists have recently 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 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.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

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 that divide the fine-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 potential drawback of 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 the underlying test score. We ask whether the label itself causes students to alter their decisions about pursuing post-secondary education.

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 post-secondary plans and their college enrollment decisions?
2: Are the college enrollment decisions of students who did not initially plan 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 for particular students?

Setting:

Description of the research location.

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.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features or characteristics.

We pool data across several years, examining students who took the 8th or 10th grade mathematics examinations in the spring of 2003 through 2006. These students are members of the graduating cohorts of 2005 through 2008. 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 repeated the grade and are taking the test for a second time. The extent to which we can examine each of these outcomes for specific cohorts depends on the timing of the initial test and outcome data collection. As seen in Table 1, each analysis uses a different number of years of data (please insert Table 1 here).

Intervention / Program / Practice:

Description of the intervention, program or practice, including details of administration and duration.

Our key "intervention" is the performance label that students earn on the state standardized test. 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 simply used to hold schools and districts accountable. However, the 10th grade examination is a high-stakes examination that students must pass to graduate from high school. As a result, in 10th grade we focus on students who fall well above the passing cutoff.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

Our analyses use the regression-discontinuity design. By examining students near each cut score, we essentially seek to estimate the 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:

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

The difference between these parameters provides an unbiased estimate of the causal impact of the classification for students at the cut score (see Lee & Lemieux, 2010, for a clear discussion).

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.

Our implementation of this regression-discontinuity strategy follows the approach laid out by Imbens and Lemieux (2008). We select a bandwidth (h) to govern the amount of smoothing in the local linear regression analysis. We choose an optimal bandwidth (h^*) for each analysis using a well-defined statistical fit criterion and a cross-validation procedure described by Imbens & Lemieux (2008).[†] We then estimate the difference between γ_{above} and γ_{below}, using local linear regression. The parameter of interest can be estimated in one step, by fitting the model presented in equation (1) using observations that fall only within one bandwidth (h^*) on either side of the relevant cut score.[‡] In its basic formulation, then, our causal estimates derive from a linear probability model of the following form:

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

for the i^{th} student. In this model, β_2 represents the causal effect of interest.

Findings / Results:

Description of the main findings with specific details.

We find strong evidence that earning a more positive performance label causes urban, low-income students to update their educational plans and to attend college at greater rates. The effects are small, but substantively important. As seen in Table 2, being classified as Needs Improvement as opposed to Warning in 8th grade increases the probability of enrolling in college by 2.1 percentage points (p=0.028).[§] We find no overall effect of labeling among students earning Advanced instead of Proficient labels in grade 8. One possible explanation is that only 3.6% of Massachusetts low-income urban students (and only 16% of all students) earn an Advanced rating. Consequently, students scoring near the Advanced/Proficient cut-score are very high-performing. However, there is a substantial response to the more positive Advanced label in grade 10, where 16% of low-income urban students (and 36% of all students) score in the highest

[†] $h^* = \arg\min_h \frac{1}{N} \sum_{i=1}^{N} (\hat{GRAD}_i(h) - GRAD_i)^2$, where $\hat{GRAD}_i(h)$ 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.

[‡] 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.

[§] Again, all p-values derive from one-tailed tests.

category. Here, earning Advanced raises the probability of enrolling in college by 5.1 percentage points (p=0.012). In general, these results suggest that the information embedded in the performance label is important to students' decisions to enroll in college. Since only 35% percent of urban, low income 8th graders enroll in college within one year of cohort graduation, a 2.1 percentage point difference represents a substantial effect (insert table 2 here).

We find that for students who reported that they plan to attend a four-year college, performance labels do not affect college plans or college enrollment decisions. In contrast, for students who reported before they took the mathematics examination that they did not plan to attend a four-year college, earning a more positive label has a substantial, positive effect across all cutoffs and outcomes. We present these results in Table 3. For example, for students who do not plan to attend a four-year college, being classified as Advanced, rather than Proficient, on the 10th grade test increases the probability that they will attend college by 9.9 percentage points (p=0.005) (insert Table 3 here).

The findings presented above clearly indicate 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 likely 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 and increase 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 may be less likely to consider themselves "college material". This would represent a negative updating of their abilities. Unfortunately, since each group represents our estimate of the counterfactual for the other, our regression-discontinuity estimates only reflect the difference between them.

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 they earned in a previous grade. However, we assume that no updating occurs if the new information matches students' priors. Interestingly, we find different patterns of responses at different parts of the test score distribution. For example, for 8th graders near the bottom of the distribution, scoring Needs Improvement instead of Warning has no effect for students who had previously scored Warning – thus, we see no evidence of an encouragement effect. However, we find substantial effects for students who had received a label of Needs Improvement in the past, which we can interpret as a discouragement effect of earning Warning instead of Needs Improvement. At the top of the distribution, the patterns are reversed, suggesting that earning a positive label encourages higher performing students. We present these results for the 8th grade test in Table 4. For the 10th grade test, nearly all students near the Advanced/Proficient had scored Proficient or lower on the 8th grade test. Thus, the encouragement effect appears to predominate at the top of the distribution (insert Table 4 here).

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

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, 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 also provoke 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 10^{th} grade.

Our findings have an important methodological implication for research that aims to identify the causal effects of policy interventions using a regression-discontinuity design. Often researchers take advantage of policies that assign students to treatment on the basis of whether their value on a continuous 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 about the intervention's effects may be confounded with the effect of 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 itself – ostensibly a fairly weak signal – has a quite powerful effect on student outcomes, including college enrollment decisions that occur several years after the test. 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 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 find ways to support those students who earn the "Warning" label.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

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

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.

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. Paper prepared for the Social Inequality and Educational Disadvantage conference, Washington, D.C.

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.

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

Appendix B. Tables and Figures
Not included in page count.

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

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

Table 2. Estimated effect of earning the more positive performance label at different cutoffs and on different outcomes, for urban, low-income students scoring near the cut point. Cell entries include the parameter estimate, standard error (in parentheses), optimal bandwidth used, sample size, and asterisks to denote inference.

	8th Grade		10th Grade
	Needs Improvement/ Warning	Advanced/Proficient	Advanced/Proficient
College attendance	0.021 *	0.006	0.051 *
	(0.009)	(0.025)	(0.020)
	h=3	h=9	h=8
	5,799	2,871	4,171

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

Table 3. Estimated effect of earning the more positive performance label at different cutoffs and on different outcomes, for urban, low-income students scoring near the cut point, by whether they express plans to attend a four-year college after high school. Cell entries include parameter estimates, standard errors (in parentheses), and asterisks to denote inference.

	Students with College Plans	Students Without College Plans	Sample Size
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
8th Grade Advanced/Proficient Cutoff			
Express 4-year college plans (grade 10)	0.007	0.137 *	2,294
	(0.023)	(0.071)	h=8
10th Grade Advanced/Proficient Cutoff			
College Attendance	0.027	0.099 **	3,316
	(0.035)	(0.033)	h=8

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

Table 4. Estimated effect of earning the more positive performance label at different cutoffs and on different outcomes, 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 asterisks to denote inference.

	Students with lower scores on prior test	Students with higher scores on prior test	Bandwidth
8th Grade Needs Improvement/Warning Cutoff			
Express 4-year college plans (grade 10)	0.002	0.023	h=5
	(0.037)	(0.025)	
	4,476	4,558	
Attend college	-0.006	0.108 ***	h=3
	(0.032)	(0.022)	
	1,531	1,077	
8th Grade Advanced/Proficient Cutoff			
Express 4-year college plans (grade 10)	0.065 ***	-0.007	h=8
	(0.015)	(0.038)	
	3,226	1,132	
Attend college	0.074 **	0.030	h=9
	(0.023)	(0.058)	
	1,151	276	

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