Overview

- Unintended consequences of test-based accountability in Massachusetts

- State divides continuous performance measure into different groups (labels/treatments)

- Design: Natural experiment with a regression discontinuity design comparing students just above and below the cut scores
Overview

• “Low-stakes” examinations
  – No official, state-defined consequences for students (e.g., 8th grade tests)
  – Students receive test score and performance label

• Does the performance label itself affect future student outcomes – attending college?

• Answer: Yes, for urban, low-income students
Motivation and Context
Responses to test performance

- Effects of performance labels could act in several ways:
  - Directly through students’ self-judgments
  - Indirectly through parents’ or teachers’ responses
Direct factors

- Students’ judgments about their potential for academic success affect educational outcomes

  "Competence is much more fragile – and malleable – than we tend to think. … How a student construes the way he or she is viewed and treated by others matters a lot" (Aaronson & Steele, 2005)

- Educational expectations predict attainment above academic achievement (Jacob & Wilder, forthcoming)
Indirect factors

- Teachers’ expectations of student performance matter for student outcomes (Jussim & Harber, 2005).
  - “Soft bigotry of low expectations”
Research questions

- Does earning a more positive performance label on the Massachusetts state mathematics test affect post-secondary educational decisions for urban, low-income students?

- Is the effect different for students with and without plans to attend a four-year college?

- Are we seeing encouragement effects or discouragement effects?
Research Design
State testing in MA

- Grades 3-8 and 10, math and ELA (MCAS)

- RD Design – Assignment to treatment based on cutoff

- Data from five test years (2002-03 to 2006-07)
Your child's performance levels and scores

<table>
<thead>
<tr>
<th>Subject</th>
<th>Performance level</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Language Arts</td>
<td>Advanced</td>
<td>274</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Advanced</td>
<td>260</td>
</tr>
<tr>
<td>Science and Technology/ Engineering</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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 grey bar shows the range of likely scores your child could have received if he or she had taken the test multiple times.

Example:

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

Performance Level

- **Advanced**: Students at this level demonstrate a comprehensive and in-depth understanding of challenging subject matter and provide sophisticated solutions to complex problems.
- **Proficient**: Students at this level demonstrate a solid understanding of challenging subject matter and solve a wide variety of problems.
- **Needs Improvement**: Students at this level demonstrate a partial understanding of subject matter and solve some simple problems.
- **Warning**: Students at this level demonstrate a minimal understanding of subject matter and do not solve simple problems.
Research Design: Analytic Approach

Main outcome:

- College attendance
- Express college-going plans in grade 10 (before they take the 10\textsuperscript{th} grade test)

Regression discontinuity design following Imbens & Lemieux (2008)

Use a nonparametric approach with local linear regression
Key Findings
Key Findings: Effect of classification

- Earning a better performance label improves outcomes for urban, low-income students
  - 10\textsuperscript{th} grade Advanced vs. Proficient:
    - Increases prob. of attending college by 5.1 pct pts (p=.012)
  - 8\textsuperscript{th} grade Needs Improvement vs. Warning:
    - Increases prob. of attending college by 2.1 pct pts (p=.029)
Key Findings: Educational expectations

- Strong evidence that the effects of classification depend on post-secondary educational plans

- Effects are much bigger for students who report before they take the test that they do not plan to attend a four-year college
Key Findings: Educational expectations

- For students without 4-year college plans:
  - Being classified as NI vs. W in 8th grade increases prob. of expressing four-year college-going plans in 10th grade by 4 % points (p=0.044)
  - Being classified as A vs. P in 8th grade increases prob. of expressing four-year college-going plans in 10th grade by 14 % points (p=0.037)
  - Being classified as A vs. P in 10th grade increases prob. of attending college by 10 % points (p=0.005)
Effect of classification for students without 4-year college plans (10th grade Adv. vs Prof.)

![Graph showing the effect of classification on the probability of attending college. The x-axis represents the 10th grade mathematics score relative to the advanced/proficient cut-off, while the y-axis represents the probability of attending college. Two lines are plotted: one for four-year college plans (red line) and one for not four-year college plans (dashed blue line). The graph illustrates the trend of higher probability for students with four-year college plans as their mathematics score increases.]
Effect of classification for students without 4-year college plans (8th grade Adv. vs Prof.)
Encouragement or Discouragement?

- Suggestive evidence using past test histories

7th Grade
- Proficient
- or
- Advanced

8th Grade (A vs. P)
- or

or
## Encouragement or Discouragement?

<table>
<thead>
<tr>
<th>Effect on probability of attending college</th>
<th>Lower scores - past test</th>
<th>Higher scores - past test</th>
</tr>
</thead>
<tbody>
<tr>
<td>8th Grade NI/W</td>
<td></td>
<td>0.108*** (0.022)</td>
</tr>
</tbody>
</table>

**Legend:**
- **Encouragement**
- **Discouragement**
Conclusion & Implications
Conclusion & Implications

- Strong evidence that labeling matters, particularly for vulnerable students
  - Competence is malleable (weak priors)
  - Labels on exams are important signals

- Question becomes: why do labels matter when students receive the test scores, too?
  - Cognitive limitations
  - Emotional responses
Conclusion & Implications

- For policymakers:
  - Must find ways to support all students, including those with lower labels
  - Need to pay particular attention to most vulnerable students
  - Discouragement may play greater role for students at bottom of test score distribution
  - Some of effect may operate through teachers – need to be careful about use of performance labels
Implications for research

- Raises questions about RD strategy to draw causal inferences about an intervention
  - If label matters, may conflate effect of labeling with effect of intervention
  - Researchers must think carefully about all the pathways through which assignment to treatment could affect outcomes
Thank You
Threats to validity: Key assumptions

- Cut scores established exogenously
  - State established cut score using complicated procedures
  - Students cannot affect status other than through test score
  - Standard tests reveal no violations of key assumptions
    - No disruption in test score density at cut score
    - No disruption in covariates at cut score
Threats to validity: Key assumptions

- Relationship between outcomes and MCAS score estimated accurately
  - Non-parametric approach
  - Most results are robust to bandwidth choice
Bandwidth choice

Estimated effect of scoring Advanced in 10th grade on college-going, for students without four-year college plans, by bandwidth
Comparison of State and NAEP standards

Understanding college aspirations

![Graph showing the probability of expressing four-year college plans by gender and ethnicity.](image-url)
Research Design: Empirical Strategy

■ Key decision is bandwidth choice ($h$)
  – Bias vs. precision tradeoff

■ Use cross-validation procedure described by Imbens & Lemieux (2008)
  – $h^*$ minimizes MSE in estimating boundary points

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

■ Assess sensitivity of results to bandwidth choice
Research Design: Empirical Strategy

- Standard RD setup:
  - MATH = Recentered math score (raw score)
  - ABOVE = \{ 1, MATH \geq c \\
 0, MATH < c \}

- Local linear regression for students within one bandwidth of cut score:

\[
Y_i = \beta_0 + \beta_1 MATH_i + \beta_2 ABOVE_i + \beta_3 ABOVE XMCAS_i + X_i' \gamma + \varepsilon_i
\]

Causal effect of more positive label for students at cutoff
Threats to validity

Density of student 10th grade MCAS scores, relative to the Adv/Prof cutoff
Threats to validity

Density of student 8th grade MCAS scores, relative to the Adv/Prof cutoff
Threats to validity

Density of student 8th grade MCAS scores, relative to the NI/W cutoff
Threats to validity: Adams Scholarship

- State offers an Adams Scholarship to students based on MCAS performance in grade 10
  - Basic criteria: Advanced in one subject, Proficient in the other, top 25% of HS class on combined test score
  - Thus, analysis of the Advanced/Proficient cutoff in grade 10 may be compromised

- Exclude students for whom the A/P distinction in mathematics decides their eligibility
## Test Score Distributions - Mathematics

<table>
<thead>
<tr>
<th></th>
<th>Grade 8</th>
<th>Grade 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced</td>
<td>13%</td>
<td>34%</td>
</tr>
<tr>
<td>Proficient</td>
<td>27%</td>
<td>27%</td>
</tr>
<tr>
<td>Needs Improvement</td>
<td>31%</td>
<td>25%</td>
</tr>
<tr>
<td>Warning/Failing</td>
<td>29%</td>
<td>13%</td>
</tr>
</tbody>
</table>
## Outcomes of urban, low-income students

<table>
<thead>
<tr>
<th></th>
<th>Plan to Attend a 4-yr College</th>
<th>Attend College</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gr. 8</td>
<td>Gr. 10</td>
</tr>
<tr>
<td>Advanced</td>
<td>85%</td>
<td>85%</td>
</tr>
<tr>
<td>Proficient</td>
<td>74%</td>
<td>71%</td>
</tr>
<tr>
<td>Needs Improvement</td>
<td>63%</td>
<td>60%</td>
</tr>
<tr>
<td>Warning/Failing</td>
<td>47%</td>
<td>46%</td>
</tr>
<tr>
<td>Educational Expectations</td>
<td>Attend College</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4-yr</td>
<td>No</td>
</tr>
<tr>
<td>Advanced</td>
<td>76%</td>
<td>52%</td>
</tr>
<tr>
<td>Proficient</td>
<td>64%</td>
<td>33%</td>
</tr>
<tr>
<td>Needs Improvement</td>
<td>50%</td>
<td>24%</td>
</tr>
<tr>
<td>Warning/Failing</td>
<td>29%</td>
<td>13%</td>
</tr>
</tbody>
</table>
Additional research

- Different tests, different answers: The stability of teacher value-added estimates across outcome measures, *AERJ in press*
- Teachers’ views on NCLB: Support for the principles, concerns about the practices, *Journal of Economic Perspectives, 2010* (with R. Murnane)
- Leading the local: Teachers union presidents chart their own course, *PJE 2009* (with S. Johnson et al.)
- *Redesigning teacher pay: A system for the next generation of educators* (with S. Johnson)
- Is PAR a good investment? Understanding the costs and benefits of teacher Peer Assistance and Review programs (with S. Johnson)
- Do teachers continue to grow with experience? Evidence of long-term career growth in the teacher labor market (with M. Kraft)