

# Beyond Exclusion: The Role of High-Stake Testing on Attendance the Day of the Test

Magdalena Bennett

*The University of Texas at Austin*

Christopher Neilson

*Yale University*

Seminario Ingeniería UC

August 8th, 2023

Nicolás Rojas

*Columbia University*

# The big picture



# Motivation

- Results from **high-stakes tests** widely used in education policy
  - E.g. funding, promotions, school closures, school choice, etc.
- **Assumption**: Standardize tests used as a proxy of school quality

Is it so?

# Motivation

Answer Sheet • Perspective

## Remember the Atlanta schools' cheating scandal? It isn't over.



By Valerie Strauss  
Staff writer

February 1, 2022 at 11:18 a.m. EST



DIVERSITY & EQUITY

TACKLING RACISM

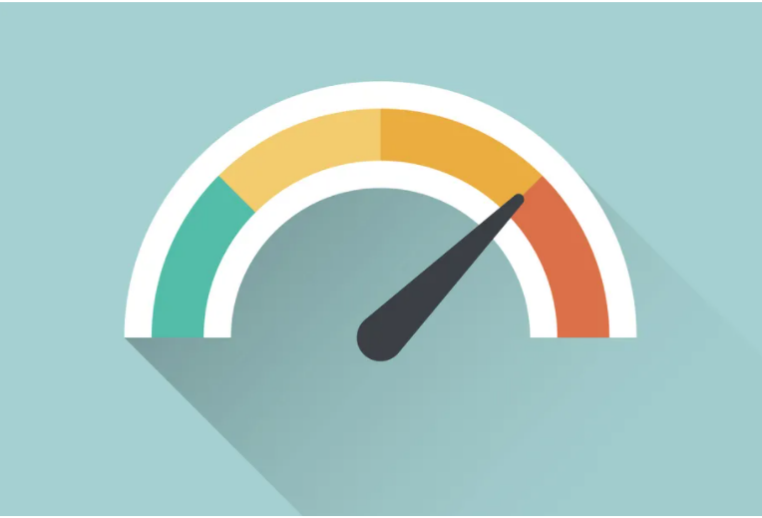
FOLLOWING PHILANTHROPY

## Looking for a home? You've seen GreatSchools ratings. Here's how they nudge families toward schools with fewer black and Hispanic students.

By Matt Barnum and Gabrielle LaMarr LeMee | Dec 5, 2019, 8:00am EST



REPUBLIC



# Motivation

- Beyond explicit cheating and socioeconomic sorting: **Students' exclusion**
  - E.g.: Reclassification of low-performers as students with disabilities (Figlio & Loeb, 2011)
  - Use of disciplinary measures to exclude low-performers (Figlio, 2006)
- Less attention on **non-representative attendance patterns**
  - Differences between scores before and after imputation (Cuesta et al., 2020)
- Schools have **incentives** to game the system
  - Especially in **high-accountability settings**

# This paper

## Attendance Patterns

- Event study approach:
  - *How do these exclusions patterns look like? Are these the same for every (type of) school and every grade?*
  - Focus beyond bottom performers
  - Robustness checks for alternative mechanisms

## Imputation Policies

- Machine learning prediction:
  - Identification of schools that are most likely gaming the system
  - Consequences of blanket policies in imputation of scores

# Outline

1. Motivation
2. Chilean educational context
3. Attendance patterns:
  - Event study for different years, grades, and performance
  - Potential mechanisms
4. Prediction approach:
  - Difference between predicted and observed distributions
  - Potential consequences of imputation
5. Conclusions and next steps

# The Chilean Educational Context



# The Chilean context: Standardized testing

- Chile has a **universal voucher system** (school choice)
- Universal standardized testing since 1980's (SIMCE)
  - For all 4th graders; then extended to other grades.
- SIMCE as **high-stake** testing:
  - Results widely available in a universal voucher system
  - Tied to teachers' bonuses
  - Tied to budget restrictions and school closures

# SIMCE and absenteeism

- Use of **pre-filled communication** for parents to be sent out by schools
  - Evidence that parents from lower-income students are less likely to receive information
- **No real consequences for low attendance:**
  - Between 2005-2007, non-representative results were marked with symbols
  - No imputation strategy so far
- Improvement of regulation for **justifying students exclusion**
  - E.g. specific disabilities (blindness) or non-Spanish speakers.

Attendance Patterns for the Day of the Test

# How to evaluate the effect of "day of the test" on abstenteeism?

- Some studies assessing the **effect of attendance manipulation**:
  - Focus on distortions (difference between imputed and observed scores) (Cuesta et al., 2020)
  - Manipulation for specific vulnerable schools (SEP) to raise scores (Feigenberg et al., 2019; Quezada & Hippel, 2017)
- **This paper**: Event study between 2011 and 2018 for all tested grades.
  - Focus on attendance by within-school performance
  - Use of alternative non-high-stake test to analyze potential mechanisms
  - Use of unpublished survey for communication and incentives around SIMCE

# Data Available

- **Standardized tests 2011-2018 (SIMCE)**
  - Scores at student and school level for different subjects (Math, Language, History, and Science)
  - Student's socioeconomic characterization (parental questionnaire)
- **Daily attendance data 2011-2018 (SIGE)**
  - Use for voucher payments (each day has ~ 2.5 million records)
- **GPA Performance 2011-2018 (Rendimiento)**
  - Use GPA performance deciles within school-grade

# Observations from our data

Data description			
Grade	Years tested	Num Schools	Num Students
2	2013, 2014, 2015	5,266	628,073
4	2011, 2013-2018	5,673	1,461,289
6	2013-2016, 2018	5,516	1,056,243
8	2011, 2013-2015, 2017	5,545	1,078,140
10	2013-2018	2,623	1,213,067

# Empirical approach for difference in attendance

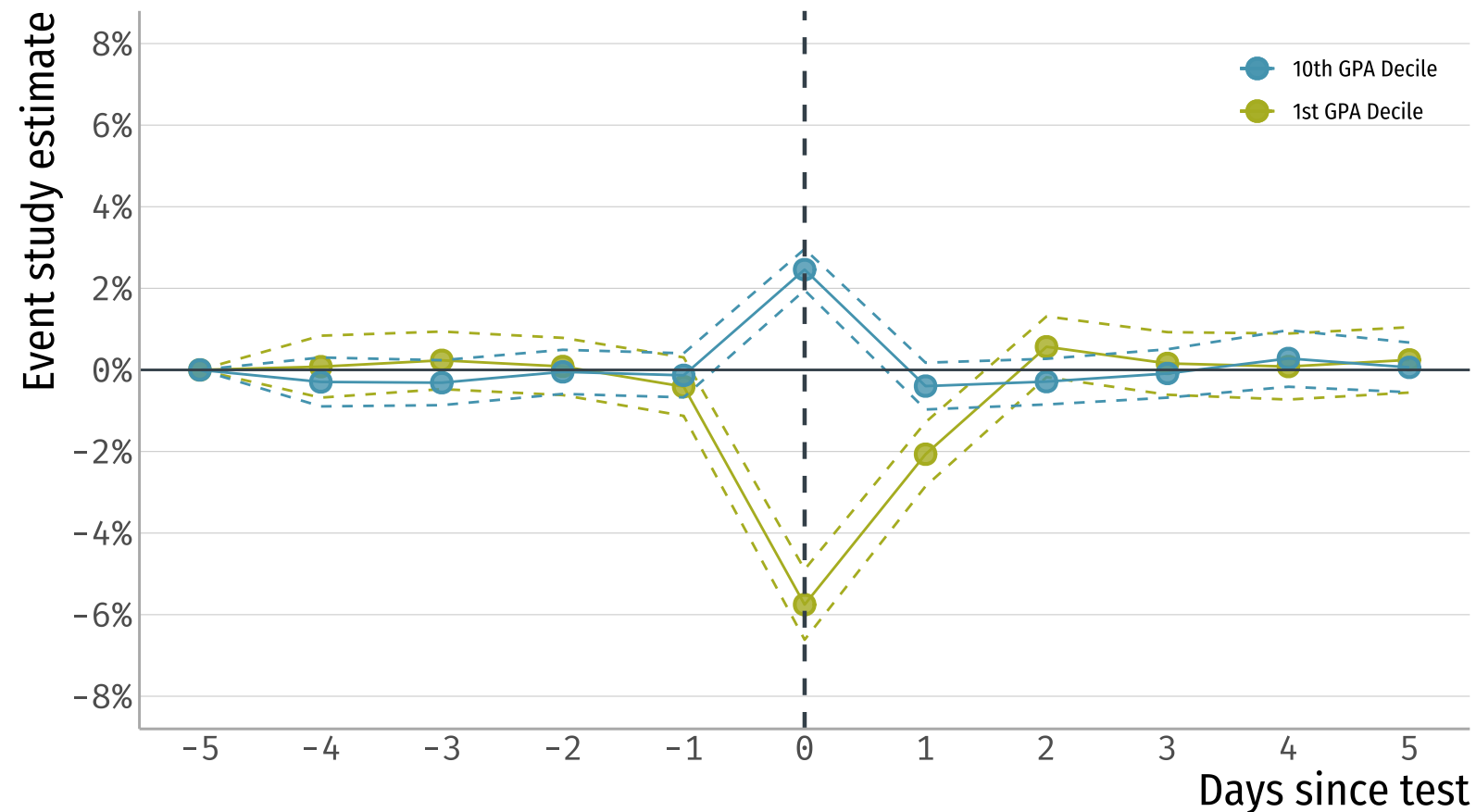
- Event study centered around the day of the test ( $T=0$ ):

$$Y_{ipsgt} = \sum_{P=1}^5 \sum_{T=-4}^5 \tau^{PT} D_{ipsgt}^{PT} + \gamma_{pt} + \alpha_i + \epsilon_{ipsgt}$$

Where

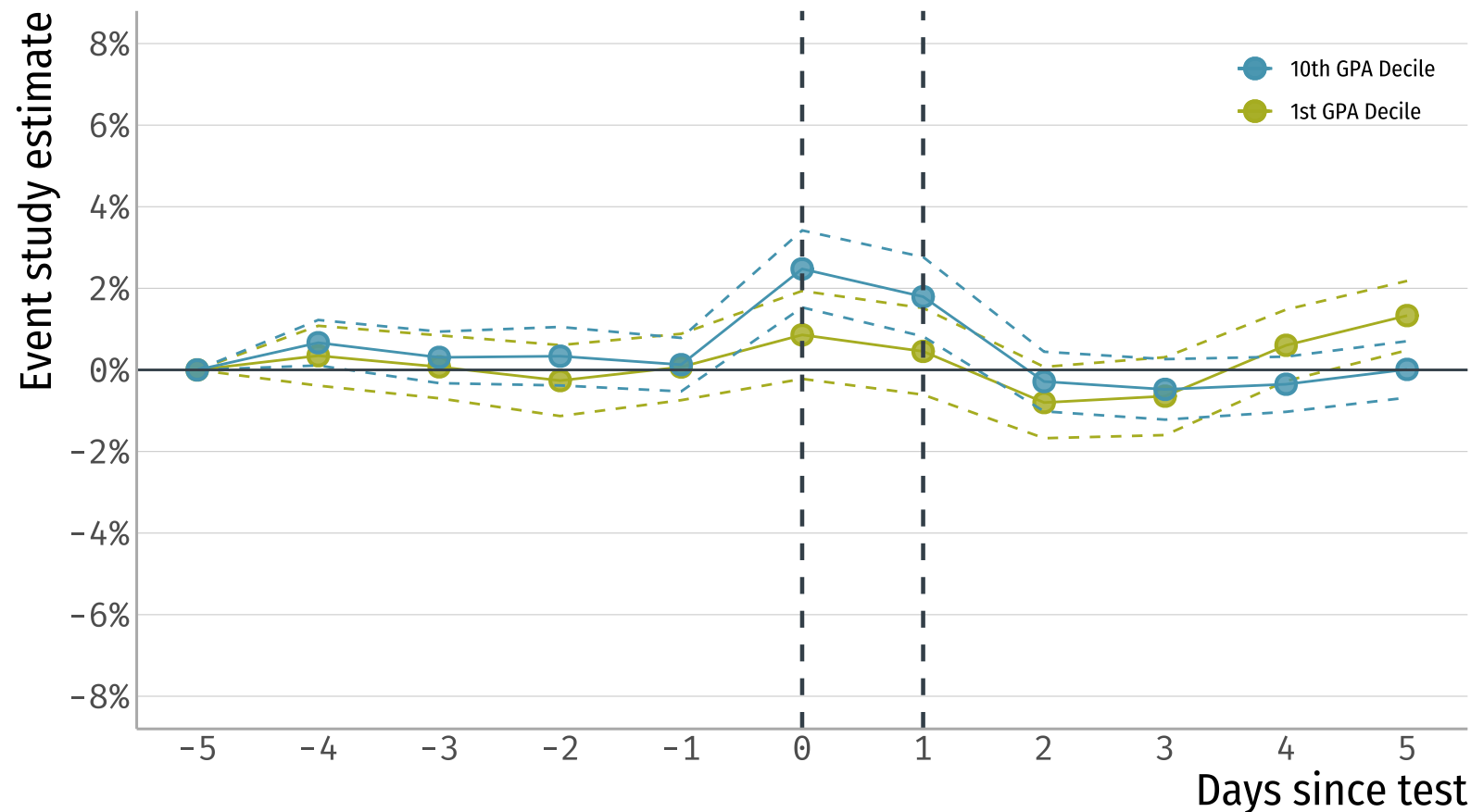
- $Y_{ipsgt}$ : Binary attendance for student  $i$ , from GPA group  $p$ , in school  $s$  and grade  $g$ , for day  $t$ .
- $D_{ipsgt}^{PT}$ : Indicator variables (lags and leads) for students that belong to a tested grade.

# Clear difference in attendance by performance for 2nd grade

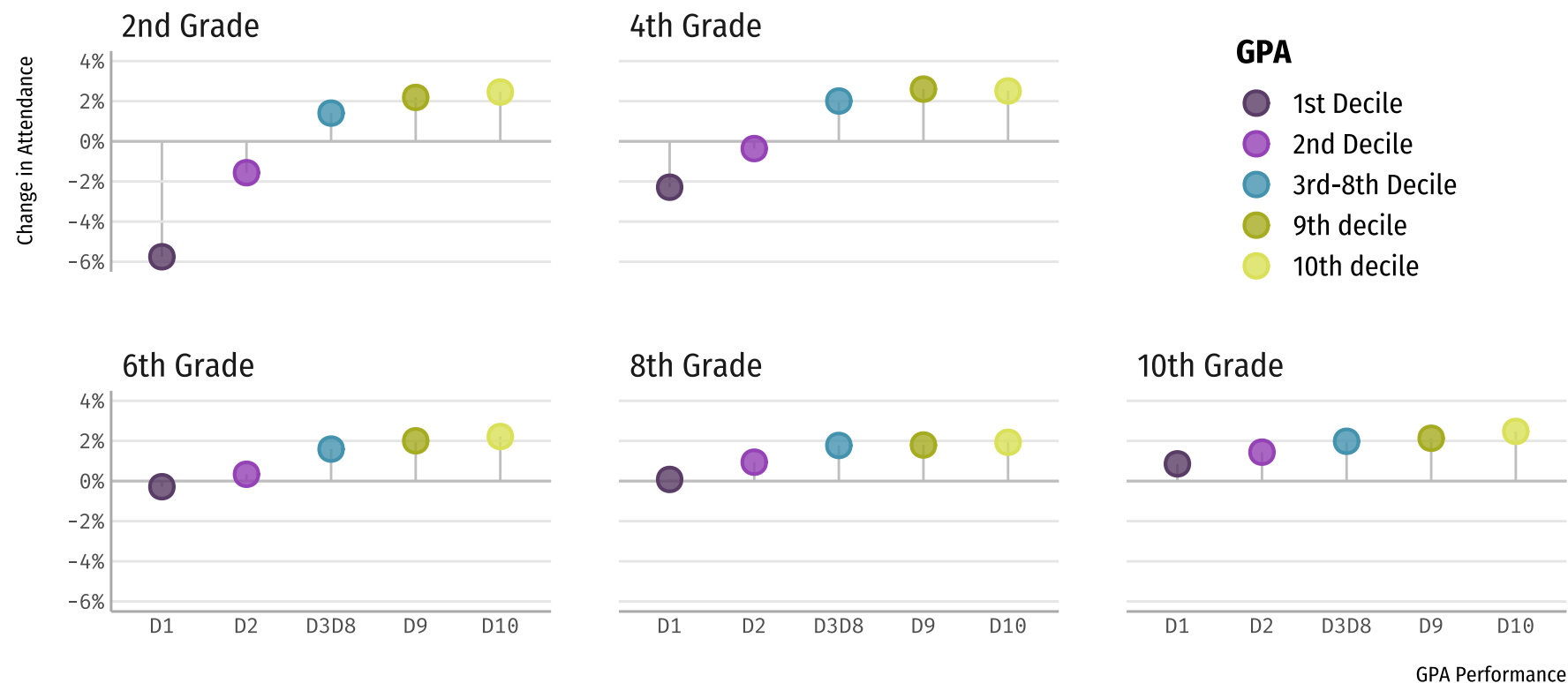




# No effect on lower performers for 10th grade



# Attendance patterns differ by grade



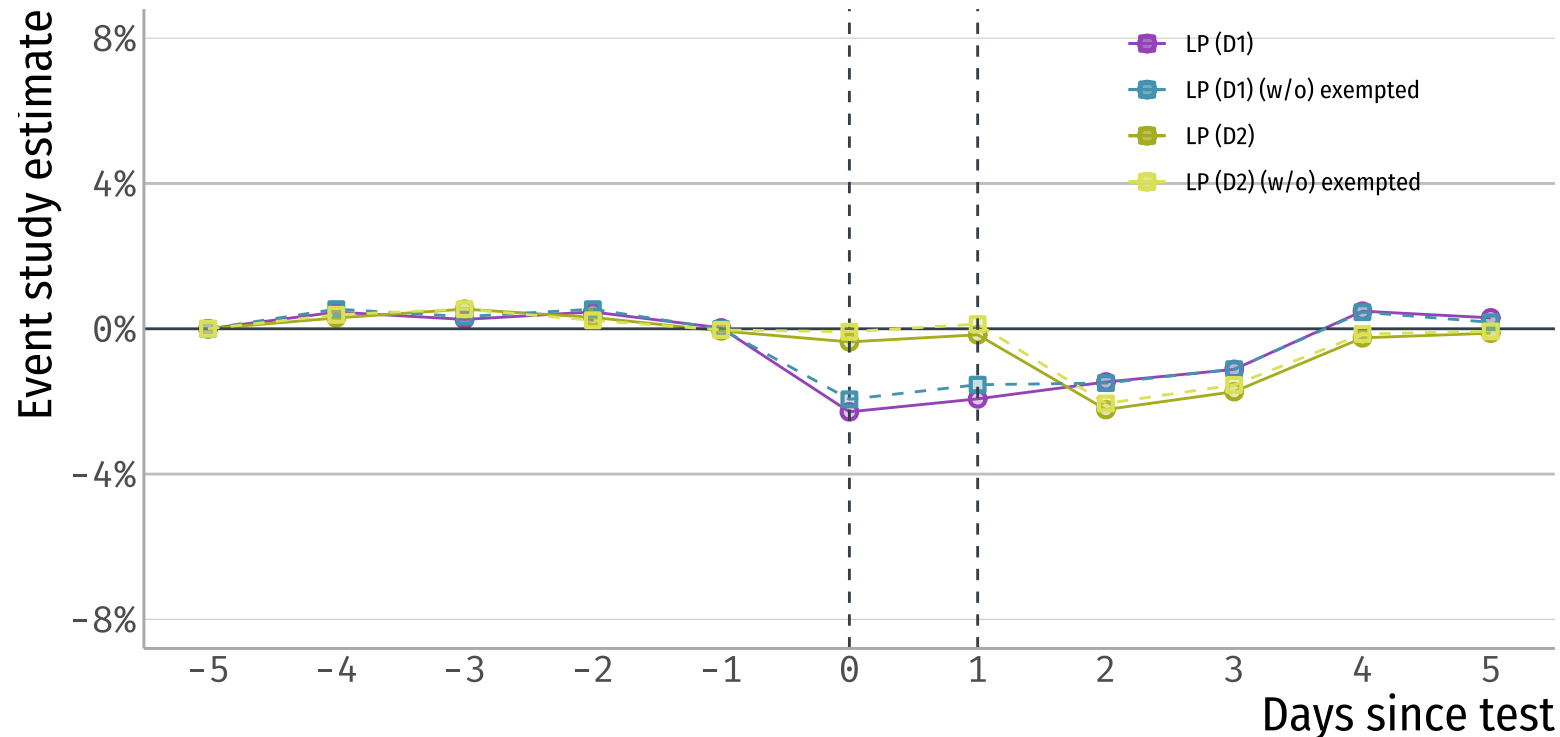
Note:  $p < 0.05$  for all estimates except those touching the 0 bar.  
Markers symbols are the coefficients of the effect of testing on attendance.

# Potential mechanisms that explain these patterns

- Students are **excluded due to other reasons** (justified)
- Students experience a **disutility from testing**
- Schools directly **(des)incentivize attendance of (lower)higher performers**

# Use of exemptions to exclude students don't tell the whole story

- Students are **excluded due to other reasons** (justified):
  - Change in exemption policy in 2012 → reduction in exempted students (flattened)
  - Results remain similar after 2012



# No evidence of self-selection from students because of testing

- Students experience a **disutility from testing**
  - Use of **no-stake test** applied to schools → No effect on attendance

Results for No-Stakes Test

Grade - Year	D1	D2	D3D8	D9	D10
2nd 2011	-0.01	0.01	0.01*	0.02	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
5th 2012	0.00	-0.01	0.00	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
6th 2011	0.02*	0.01	0.01**	0.01	0.00
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)
6th 2017	0.00	0.03	0.01	0.01	0.00
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)
11th 2012	0.00	0.00	0.00	-0.02**	0.00
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)

# Differences in communication and incentives between high and low performers

- Schools directly (des)incentivize attendance of (lower)higher performers
  - 2017 survey for students in test-taking grades.

Results for 4th Grade				
GPA Decile	Told	Notification	Preparation	Grades
D1	-0.06***	-0.11***	-0.08***	0.14***
	(0.00)	(0.00)	(0.00)	(0.00)
D10	0.06***	0.05***	0.05***	-0.2***
	(0.00)	(0.00)	(0.00)	(0.00)
Baseline	0.89***	0.87***	0.89***	0.39***
	(0.00)	(0.00)	(0.00)	(0.00)

# Differences in communication and incentives between high and low performers

- Schools directly (des)incentivize attendance of (lower)higher performers
  - 2017 survey for students in test-taking grades.

Results for 10th Grade				
GPA Decile	Told	Notification	Preparation	Grades
D1	-0.02***	-0.01***	-0.02***	0.05***
	(0.00)	(0.00)	(0.00)	(0.00)
D10	0.01***	0.00	0.00	-0.03***
	(0.00)	(0.00)	(0.00)	(0.00)
Baseline	0.95***	0.78***	0.82***	0.33***
	(0.00)	(0.00)	(0.00)	(0.00)

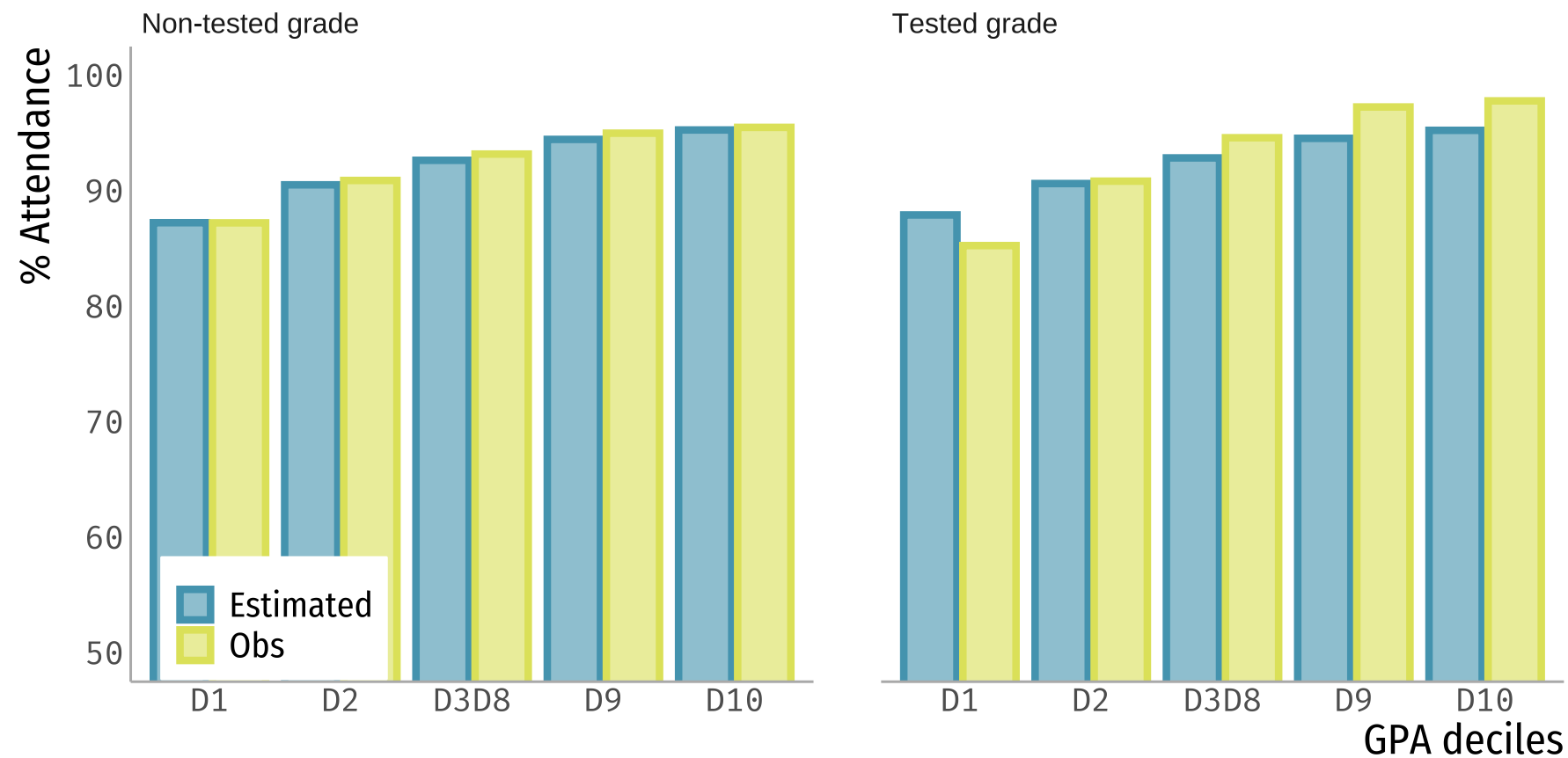
# Predicting the Counterfactual



# How do these results compare to predicted counterfactual?

- Can we use **this existing rich panel data** to predict attendance on the day of the test *as if it was a regular day*?
- Use **GPBoost** with panel data for **attendance prediction**
  - Combines Gaussian Processes and Gradient Boosting.
  - Model includes random effects for both subjects and time.
  - Predictive variables include day of the week, grade, GPA group, and sibling's attendance.
- Use data for 4th grade (2017):
  - Data **before** the test to **predict attendance on the day of the test**.

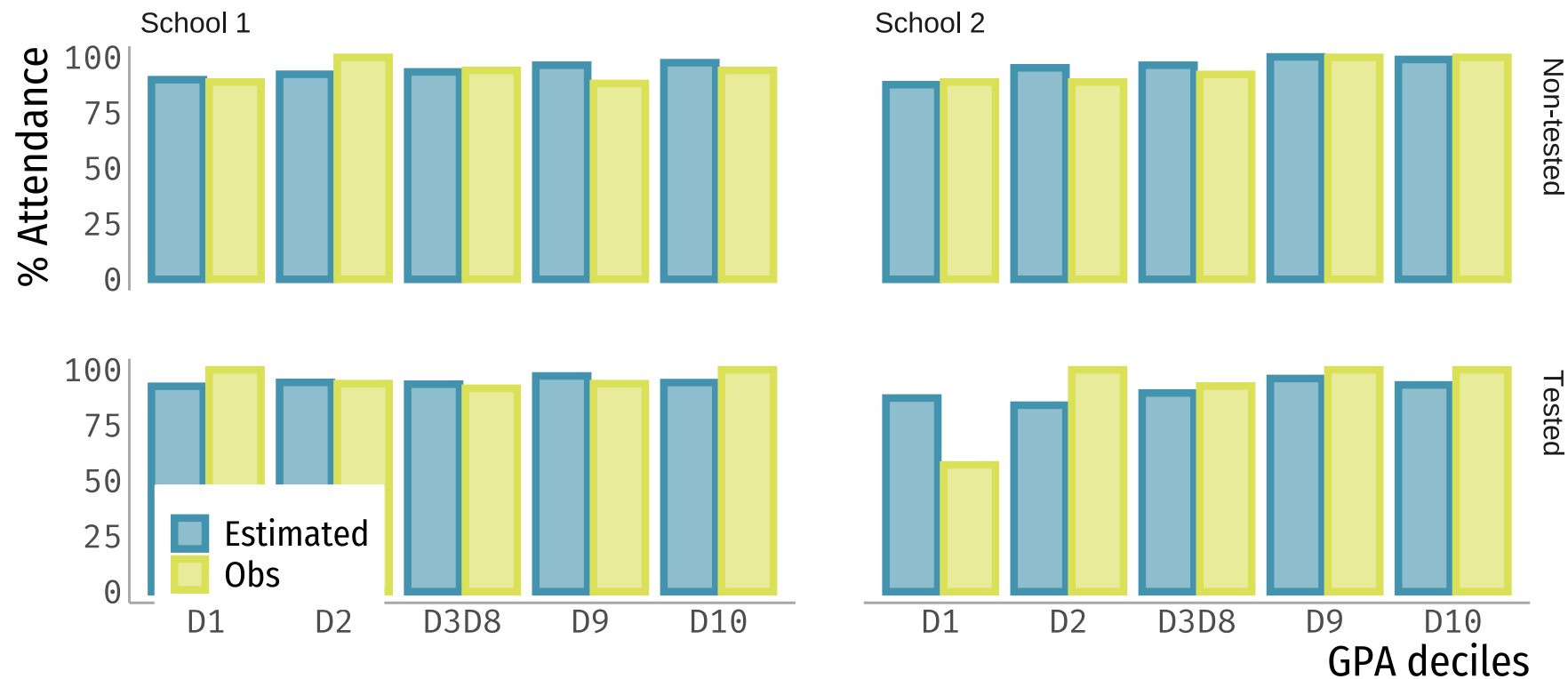
# Overall predictions over performance distribution



# Example: Comparisons between schools?

**School1**  
Math: 258  
Language: 260

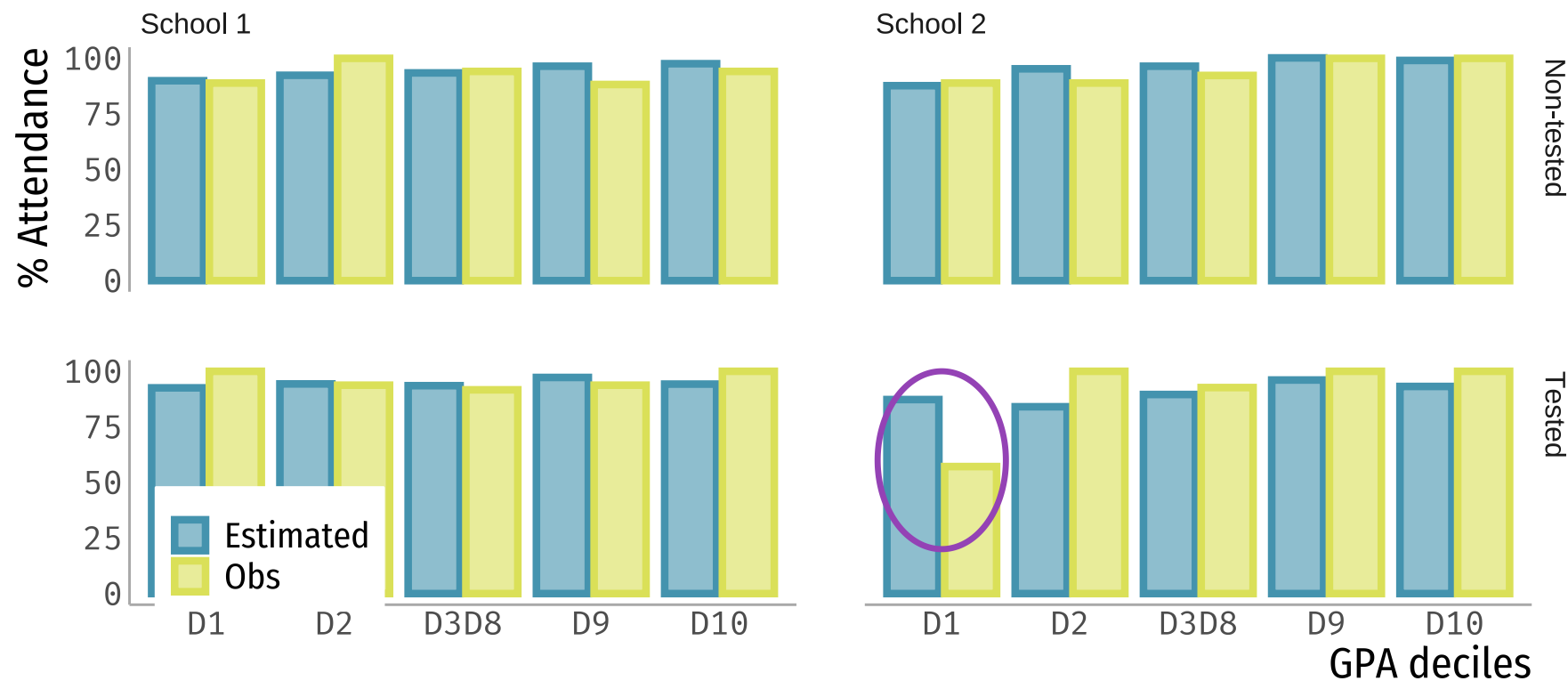
**School2**  
Math: 259  
Language: 256



# Example: Comparisons between schools?

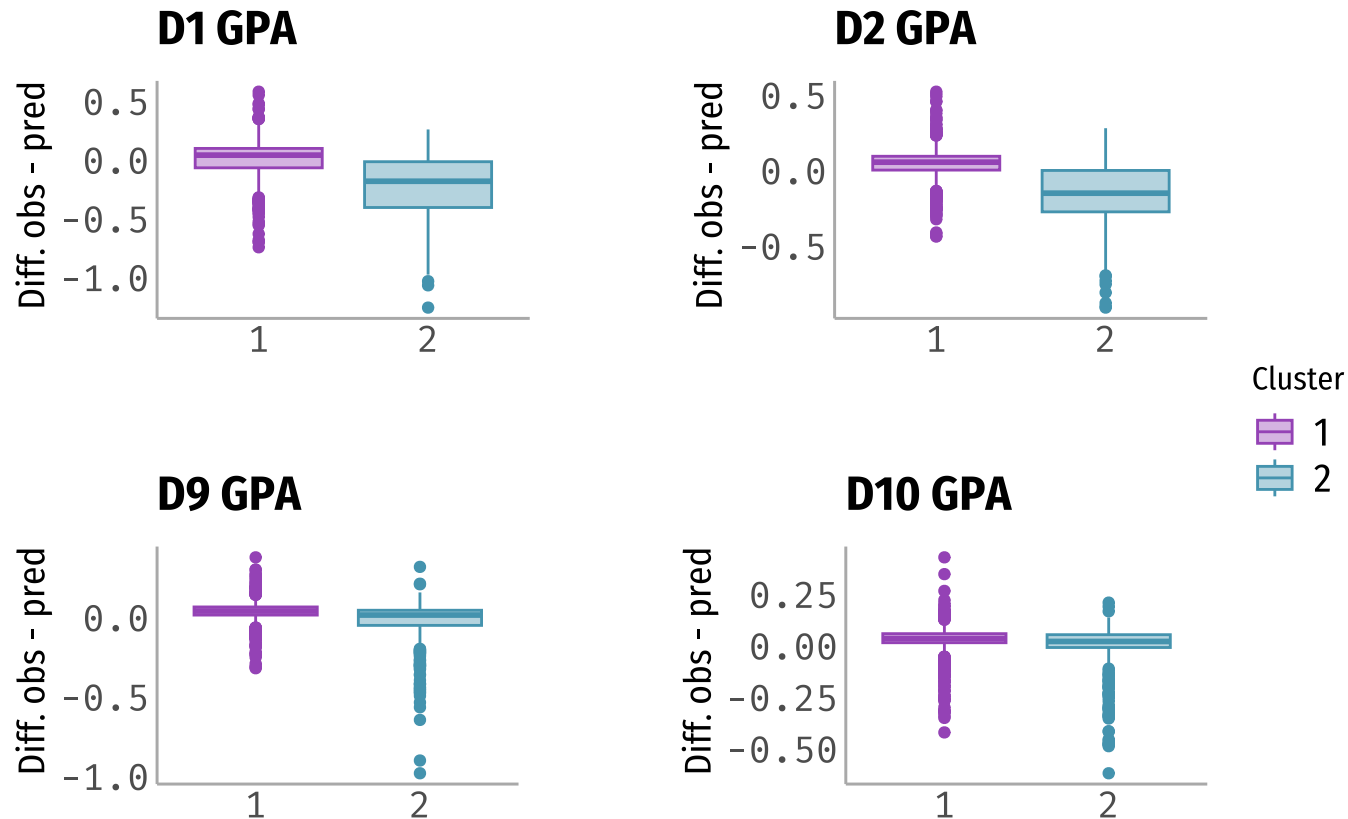
**School1**  
Math: 258  
Language: 260

**School2**  
Math: 259  
Language: 256



# Can we characterize these schools?

- **K-means clustering** Use differences between predicted and observed attendance.
  - 2 optimal clusters



# Schools that appear to exclude lower-performing students are also more vulnerable

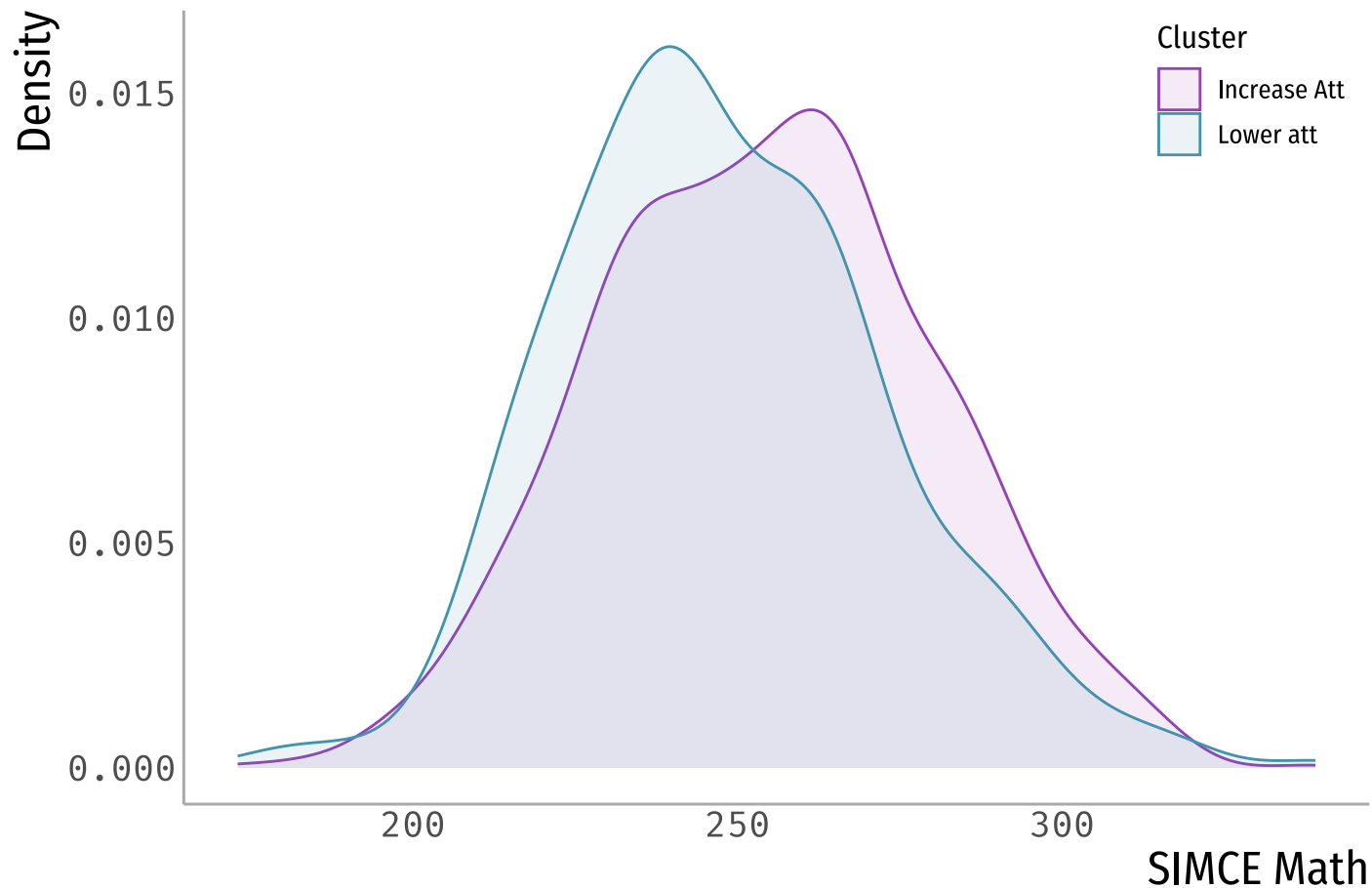
	Cluster 1 Increase att (N=1094)		Cluster 2 Lower att (bottom) (N=346)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
Avg. SIMCE Lang	258.84	22.38	252.62	23.84	-6.22	0.00
Avg. SIMCE Math	254.42	25.70	247.80	25.15	-6.62	0.00
Public	0.35	0.48	0.42	0.49	0.07	0.03
SEP status	0.84	0.37	0.88	0.33	0.03	0.11
% Priority Students	0.48	0.19	0.52	0.19	0.04	0.00
Diff D1 GPA	0.02	0.15	-0.22	0.27	-0.24	0.00
Diff D2 GPA	0.05	0.11	-0.17	0.21	-0.22	0.00
Diff D9 GPA	0.04	0.06	-0.03	0.15	-0.07	0.00
Diff D10 GPA	0.03	0.07	-0.01	0.12	-0.04	0.00

Note: Diff DX GPA represents the difference between obs. attendance and predicted attendance for decile X

# Implications for imputation policies

- **How to handle this absenteeism problem?**
  - E.g.: Observed attendance (no imputation), attendance as if the test hadn't happened (impute "typical day"), everybody is present.
- Proposals to impute **lowest scores for absent students** to disincentivize arbitrary exclusion
  - Most vulnerable schools have higher absenteeism rates → Increase inequality and non-representativeness

# There are differences in score distributions between clusters





# Differences in scores and attendance

- The previous differences between types of schools **does not capture the true difference** given non-representative attendance patterns.
- Two incentives working simultaneously:
  - Incentive for **lower-performing students not to attend** the day of the test
  - Incentive for **higher-performing students to attend** the day of the test
- We will focus on solving the **first one**.

# Some imputation exercises

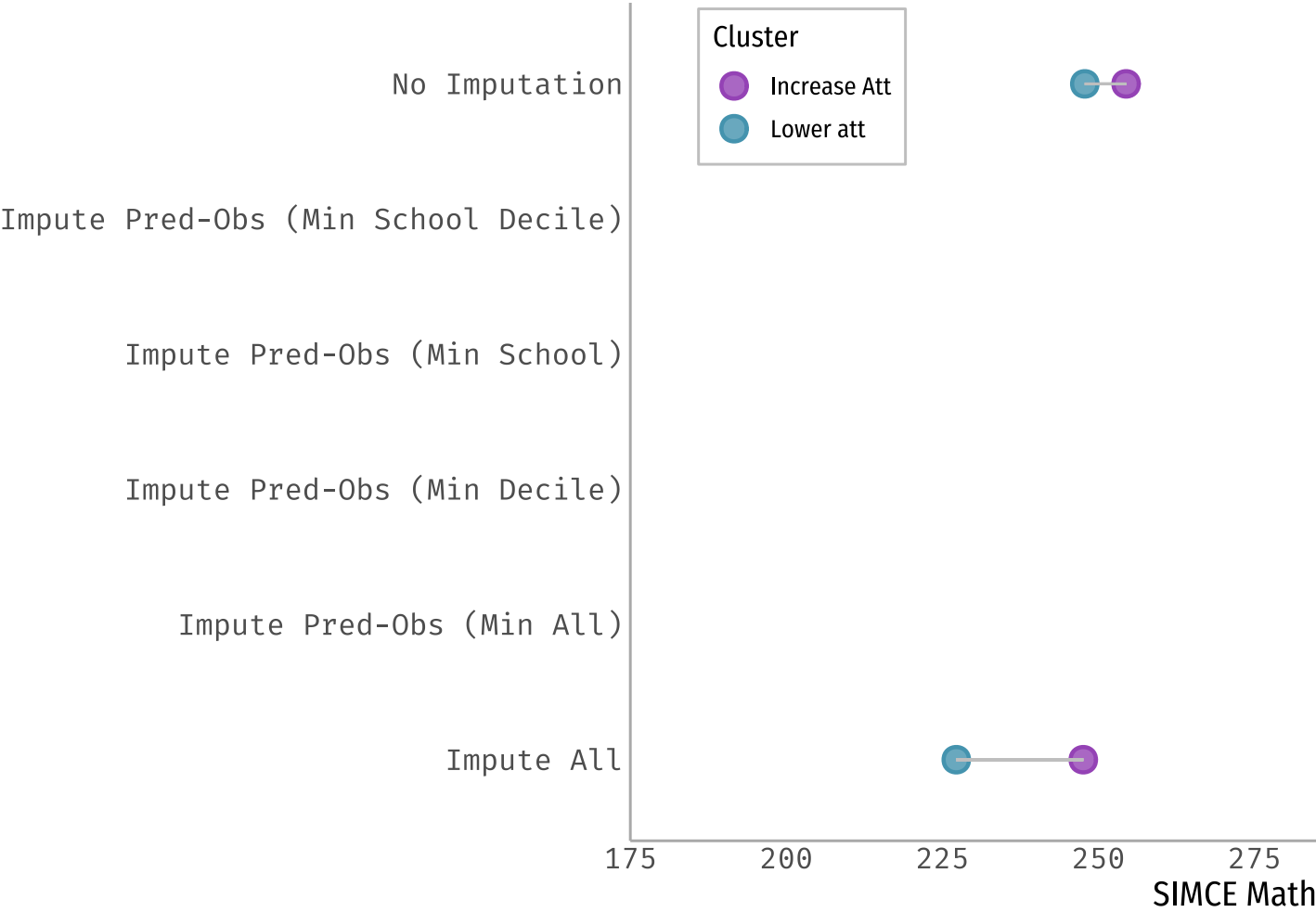
How can we **impute missing scores**?

- **Scenario 1**: Not impute at all. Show observed distributions.
- **Scenario 2**: Impute by decile only for the difference between predicted and observed attendance.
  - Imputed score: (a) overall min, (b) decile min, (c) min school, or (d) min decile by school.
- **Scenario 3**: Impute every missing student.
  - Imputed score: overall min

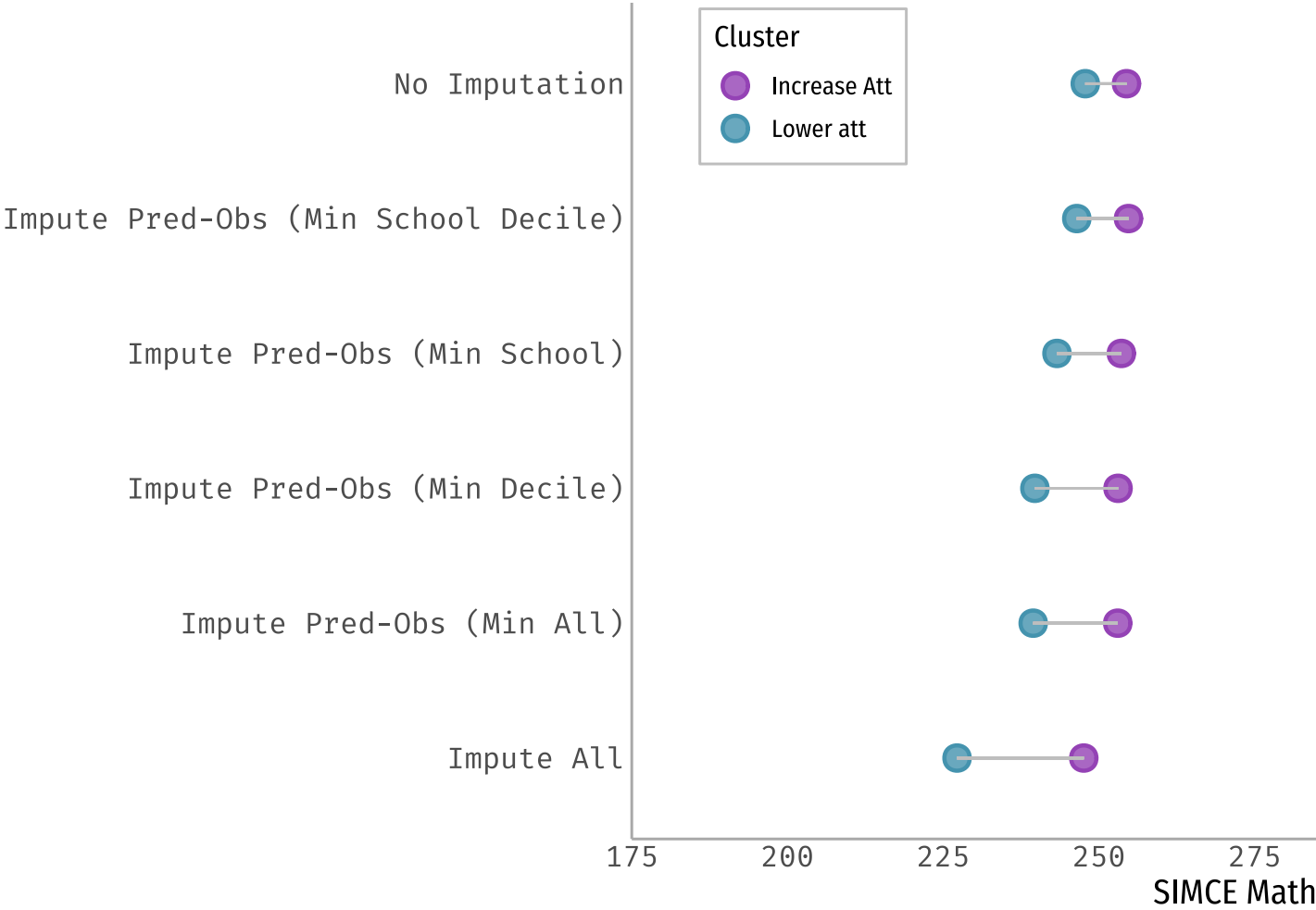
## **Some caveats:**

- Difference between predicted and obs. captures total incentives/disincentives in attendance.
- Imputed score might be too optimistic (e.g. real score would be lower than observed distribution)

# Scenario 1 vs Scenario 3: No imputation and Impute all



# Imputing Predicted - Observed is less extreme



Let's Wrap Up...

# Conclusions and next steps

- Non-representative patterns of absenteeism **beyond exclusion of low-performers**
  - High heterogeneity between schools
- **Communication strategies** play important role for **lower-performing students**
- Impact of **imputation policies**?
  - Work in progress: How does non-representativeness and different imputation strategies impact policies and information provision? What score do we impute and for whom?
- Importance of **data availability**

# Beyond Exclusion: The Role of High-Stake Testing on Attendance the Day of the Test

Magdalena Bennett

*The University of Texas at Austin*

Christopher Neilson

*Yale University*

Seminario Ingeniería UC

August 8th, 2023

Nicolás Rojas

*Columbia University*